Proceedings of the 15th Conference on

Theoretical Aspects of Rationality and Knowledge

TARK 2015

Editor

R. Ramanujam

The Institute of Mathematical Sciences, Chennai, India

ISBN number 978-0-692-46138-9

Chairs

Program Chair: R. Ramanujam (IMSc, Chennai)
Local Organizing Chair: Kevin T. Kelly (CMU)
Conference Chair: Joseph Y. Halpern (Cornell)

Program Committee

Eleonora Cresto
Clare Dixon
Edith Elkind
Amanda Friedenberg
Sujata Ghosh
Andreas Herzig
Bettina Klaus
Kevin Kelly
Yoram Moses
Andrés Perea
Sophie Pinchinat
R. Ramanujam (chair)
Francesca Rossi
Olivier Roy
Burkhard Schipper
Hans van Ditmarsch
Yanjing Wang
Michael Wooldridge

Local Organizing Committee

David Danks
Teddy Seidenfeld
Mary Grace Joseph
Jacqueline Defazio
Rosemarie Commisso

Conference Program

Invited Papers

Foreword

This volume consists of papers presented at the **Fifteenth conference on Theoretical Aspects of Rationality and Knowledge (TARK)** held at *Carnegie Mellon University, Pittsburgh, USA* from June 4 to 6, 2015.

It has been my pleasure to be part of the TARK community since the first conference at Asilomar, California, in 1986, principally due to the encouragement of Rohit Parikh, one of the founders of TARK. This conference is uniquely situated as one that brings together researchers from a wide variety of fields, including Artificial Intelligence, Cryptography, Distributed Computing, Economics and Game Theory, Linguistics, Philosophy, and Psychology. It has played an important role in our understanding of interdisciplinary issues involving reasoning about rationality and knowledge.

This year we had 63 submissions out of which 18 were accepted as contributed talks and 9 as poster presentations for the programme. I am very grateful for having the cooperation and advice of 17 other members of the multidisciplinary program committee: Eleonora Cresto (CONICET, University of Buenos Aires, Argentina), Clare Dixon (University of Liverpool, UK), Edith Elkind (Oxford University, UK), Amanda Friedenberg (Arizona State University, USA), Sujata Ghosh (Indian Statistical Institute, India), Andreas Herzig (IRIT, Toulouse, France), Bettina Klaus (University of Lausanne, Switerland), Kevin Kelly (Carnegie Mellon University, USA), Yoram Moses (Technion, Tel Aviv, Israel), Andrés Perea (Maastricht University, The Netherlands), Sophie Pinchinat (IRISA, Rennes, France), Francesca Rossi (University of Padova, Italy), Olivier Roy (Bayreuth University, Germany), Burkhard Schipper (University of California at Davis, USA), Hans van Ditmarsch (LORIA, France), Yanjing Wang (Peking University, China) and Michael Wooldridge (Oxford, UK). I thank them for their hard work in providing careful reviews and for the detailed discussions about the submissions. When papers are read across disciplines, there can be keen differences in what is considered good and important; I thank the committee members for trying their best to listen to other viewpoints.

We have four eminent invited speakers in this year's TARK: Robin Clark of the University of Pennsylvania, USA, bringing a viewpoint from Psychology and Cognition to epistemic reasoning; Simon Huttegger, of the University of California, Irvine, USA, a philosopher's look at observational process and inductive logic; Sarit Kraus of Bar-Ilan University, Israel, on desiging computational agents for interacting with people, based on insights from game theory and logic; Marciano Siniscalchi, Northwestern University, USA, on foundations of rationality in sequential games. In addition, we have a tutorial on causal inference and causal discovery, jointly by Peter Spirtes of Carnegie Mellon University, USA, and Kun Zhang of Max Planck Institute for Intelligent Systems, Tübingen, Germany.

The organizing team at Carnegie Mellon University has been doing excellent work for putting everything in place for the conference, and I thank them for all the hard work. I am extremely grateful to Kevin Kelly, the chair of the organizing committee, for his terrific coordination job.

I thank the Easychair conference system for providing this important service, easing the Programme committee's job truly easy. I thank my colleagues Vaishnavi Sundararajan and S. P. Suresh of the Chennai Mathematical Institute for help with the Proceedings volume, and Anantha Padmanabha of my Institute for help with the conference web page. I thank the Institute of Mathematical Sciences, Chennai, to which I belong, for hosting the conference page and acting as publisher.

Finally, I thank Joe Halpern, for his comforting presence and guidance all along, providing inspiration to TARK.

R. Ramanujam

Institute of Mathematical Sciences, Chennai, India
Programme Chair, TARK 2015

Quine's Topiary:
Coordination and Change in an Artificial Society

Robin Clark
University of Pennsylvania Philadelphia, USA
rclark@sas.upenn.edu

ABSTRACT

This talk reports the results of a large Agent-Based model of phonetic variation; each agent in the society has its own unique representation of the signal space, yet the agents are able to coordinate their signaling behavior. We show a number of results: First, agents in a segregated but egalitarian society will blend their signals overtime if they signal to each other; agents in a segregated, but bigoted, society will maintain stable variation. Second, if the artificial society contain high status leaders—that is, the society is not egalitarian—then the signal space will actually move apart, creating variation where none existed before. We will analyze the source of this variation and show that it is a potential source of language variation and language change. Finally, we will discuss the relationship between private knowledge and social convention.

General Terms

Coordination

Keywords

Signaling behavior, artificial society, private knowledge

The Problem of Analogical Inference in Inductive Logic

[Extended Abstract]

Simon M. Huttegger
Department of Logic and Philosophy of Science
University of California, Irvine
Social Science Plaza A
Irvine, CA-92697, USA
shuttegg@uci.edu

ABSTRACT

We consider one problem that was largely left open by Rudolf Carnap in his work on inductive logic, the problem of analogical inference. After discussing some previous attempts to solve this problem, we propose a new solution that is based on the ideas of Bruno de Finetti on probabilistic symmetries. We explain how our new inductive logic can be developed within the Carnapian paradigm of inductive logic—deriving an inductive rule from a set of simple postulates about the observational process—and discuss some of its properties.

Keywords

Inductive logic; Carnap; de Finetti; Analogy; Partial Exchangeability

1. INTRODUCTION

The logical empiricist movement is often associated with using deductive logic to understand scientific reasoning. But Rudolf Carnap actually favored an inductive approach, starting with his work on inductive logic in the 1940s. Carnapian inductive logic can be thought of as a branch of probability theory that is especially concerned with predictive probabilities—the probability of future observations given past observations. Carnap spent much of the last thirty years of his life on developing an inductive logic, but even in his posthumously published works he considered the subject to be wide open to further investigations. The open problem that I wish to consider in this paper is the problem of analogical inference, which hasn't received a satisfactory answer in Carnap's original system. I shall review some of the attempts to develop an analogical inductive logic in §4. In order to set the stage, I briefly describe Carnap's program in §2 and point to its connections with de Finetti's theory of inductive inference in §3. Considering de Finetti is particularly important since he provides an alternative route to analogical inference. In §5 I discuss an especially interesting probabilistic symmetry that allows for a certain form of analogical inference. Finally, in §6 I introduce a new analogical inductive logic based on that symmetry and discuss some of its properties.

2. CARNAP'S PROGRAM

Carnap's program for developing an inductive logic as described in his 'Logical Foundations of Probability' [2] was brought to a tentative conclusion in the posthumously published 'A Basic System of Inductive Logic' [4, 5]. Carnapian inductive logic aims at finding rational foundations for the kind of inductive inferences that are used in scientific investigations. The classic example of such an inference in the tradition of Bayes and Laplace is the predictive probability of events, such as future coin flips based on past observations of coin flips. Carnap viewed all inductive inference problems as being essentially reducible to this type of inference [2].[1]

Of particular importance for Carnap are predictive probabilities based on the relative frequencies of events. For example, after observing a number of throws of a die, the predictive probability of observing a six with the next throw usually is judged to be approximately equal to the relative frequency of sixes. In his systems of inductive logic, Carnap tries to explicate the foundations of this simplest kind of inductive inference.

Independently of Carnap's program, a similar approach was developed more than two decades earlier by the Cambridge logician W. E. Johnson [21, 22]. Johnson's main contribution was only published posthumously and contained a number of gaps, which were closed by Sandy Zabell [33], who also generalized Johnson's approach to a theory that is essentially equivalent to Carnap's basic system of inductive logic. I'm going to follow Zabell's elegant treatment because it ties in neatly with the work of Bruno de Finetti (see the next section).[2]

The basic postulate in this theory of inductive inference is a symmetry requirement known as 'exchangeability' (called the 'permutation postulate' by Johnson). Suppose that there is a finite sequence of random variables $X_1, \ldots, X_n$ representing observations (e.g. coin flips), and let their probability law be $\mathbb{P}$. Like Carnap, we assume that the random variables can take on only a finite number of values. Then $\mathbb{P}$ is *exchangeable* if it is invariant under permutations of outcomes; that is,

$$\mathbb{P}[X_1 = x_1, \ldots, X_n = x_n] = \mathbb{P}[X_1 = x_{\sigma(1)}, \ldots, X_n = x_{\sigma(n)}]$$

for every permutation σ of $\{1, \ldots, n\}$. This allows us to define exchangeable probabilities of infinite sequences $X_1, X_2, \ldots$ as those for which every finite initial sequence is exchangeable. For simplicity, the sequence of random variables is often called exchangeable without referring to its probability law.

Both Johnson and Carnap use a requirement for predic-

[1] See [36] for an excellent overview for the development of Carnapian inductive logic.

[2] Kuipers [24] gives an overview of the mathematical aspects of Carnap's theory.

tive probabilities that is often called ‘Johnson’s sufficientness postulate’. This postulate says that predictive probabilities for i basically only depend on the past relative frequency of i; i.e., there is a function f such that

$$\mathbb{P}[X_{n+1} = i | X_1, \ldots, X_n] = f_i(n_i, n). \qquad (1)$$

Johnson’s sufficientness postulate judges information about types other than i to be irrelevant for the predictive probability of i—a point that is going to be important for the problem of analogical inference.

Finally, in order for conditional probabilities to be well defined, a regularity postulate is assumed to the effect that each finite initial sequence of outcomes has positive probability. It is then possible to show that the predictive probability of any outcome is equal to its relative frequency modulo some prior parameters. More specifically, if trials are not independent, then there exist parameters α_j for each outcome j such that for all n and i

$$\mathbb{P}[X_{n+1} = i | X_1, \ldots, X_n] = \frac{n_i + \alpha_i}{n + \sum_j \alpha_j}. \qquad (2)$$

(If trials are independent, then there is no learning from experience.) Here n_i is the number of times outcome i is observed in the first n trials. The parameters α_j are either all positive or all negative; they must be positive if the sequence of observations is infinite exchangeable (see [33] for a thorough discussion). The rule given by (2) is called a ‘generalized rule of succession’ (after Laplace’s special ‘rule of succession’). A generalized rule of succession expresses a mode of learning from experience. Experiences are given by past observations of outcomes, and past observations lead to predictive probabilities for future outcomes.

The inductive logic given in (2) is equivalent to Carnap’s mature basic system of inductive logic, also known as the ‘$\lambda - \gamma$-continuum of inductive methods’. The system championed in his 1950 book is much more restricted [2]. It requires that all $\alpha_j = 1$, meaning that all outcomes are judged to be equally probable prior to any observations. In his later ‘A Continuum of Inductive Methods’ [3], Carnap generalized this restricted system to one with a weight λ which regulates the effect of the equally probable prior weights. The basic system (2) extends this to arbitrary prior weights.

Especially in his early work on inductive logic, Carnap thought of symmetry principles such as exchangeability as requirements of rationality. The idea—familiar from justifications for Laplace’s principle of indifference—is that certain probabilistic symmetries should hold whenever one does not have any knowledge about the relevant underlying structure. For instance, in the absence of any evidence concerning the order of outcomes you should assume exchangeability. We will see that interpreting symmetry principles in this way puts significant constraints on how to include analogy effects into inductive logic, while the approach discussed in the next section allows for a greater variety of inductive logics.

3. DE FINETTI’S PROGRAM

Bruno de Finetti is famous for his foundational work on probability theory and inductive inference. The latter is of special importance to us here. The most fundamental result in this arena is de Finetti’s representation theorem for exchangeable sequences [9]. Exchangeability is important because it captures one of the classic situations of statistics—i.i.d. trials with unknown parameters. This is what is shown by de Finetti’s representation theorem. Suppose, for example, that X_i records whether the ith toss of a coin flip came up heads or tails, and that the infinite sequence $X_1, X_2, \ldots$ is exchangeable. de Finetti proved that this is equivalent to the probability of finite sequences of heads and tails being a mixture of i.i.d. binomial trials with unknown bias of the coin:[3]

$$\mathbb{P}[X_1 = x_1, \ldots, X_n = x_n] = \int_0^1 p^h (1-p)^{n-h} d\mu(p) \qquad (3)$$

(Here, p is the bias for heads, μ is a uniquely determined prior over biases and h is the number of heads in the first n trials.) This theorem has profound consequences for the philosophy of probability and for inductive inference [34]. Specifically, if the prior μ in the representation is a Beta distribution (or, in the more gneral case of finitely many types of outcomes, a Dirichlet distribution), then

$$\mathbb{P}[X_{n+1} = i | X_1, \ldots, X_n] = \frac{n_i + \alpha_i}{n + \sum_j \alpha_j},$$

where α_i, α_j are nonnegative parameters determining the Dirichlet distribution. This is equivalent to the Carnapian inductive logic given in (2). One difference between the two approaches lies in the underlying axiomatic foundations. In de Finetti’s case, it is given by (i) the assumption of exchangeability and (ii) the assumption that the mixing prior in the representation μ is a Dirichlet distribution. In the Johnson-Carnap approach there is no appeal to the de Finetti representation.

There is also an important interpretive issue that separates the early work of Carnap from de Finetti’s probabilistic epistemology (in his later work Carnap is closer to de Finetti’s views). de Finetti did not view exchangeability or other symmetry requirements as postulates of rationality. According to him, exchangeability is a personal judgement of an epistemic agent as to the basic structure of a learning situation. Such a judgement does not arise from the lack of knowledge but presupposes knowledge about an epistemic situation.

This view of symmetry assumptions has two important consequences, one epistemological and one formal. In the first place, for de Finetti and his followers the justification of generalized rules of succession is only a relative one. An agent should make inductive inferences provided that she assumes certain underlying symmetries about the learning situation. This is unlike the objective Bayesian tradition—which includes Bayes, Laplace, Keynes, the early work on inductive logic by Carnap, and others—where symmetry assumptions themselves are viewed not just as assumptions that one may make, but as principles every rational agent has to make under certain conditions.

de Finetti’s probabilistic epistemology is thus distinctly non-foundationalist. There is no bedrock of initial epistemic judgements that would endow all their consequences with full rationality because they are themselves requirements of rationality. For de Finetti, rationality is instead to be found in the interplay of *inductive assumptions*, such as Johnson’s sufficientness postulate or exchangeability, and rules for learning from observations. If you use such an

[3] For finite forms of this result, see [13].

inductive rule but deny its underlying assumptions, you are simply inconsistent. So, de Finetti requires a kind of relative rationality: learning from experience should be compatible with those inductive assumptions that are judged to be true.

The second consequence of de Finetti's view of symmetry assumptions lifts constraints from inductive logic. If assumptions such as exchangeability are not thought of as requirements of rationality but as personal judgements, then one might consider other kinds of symmetries whenever exchangeability does not seem appropriate. This led de Finetti to study 'partial exchangeability' [10, 11, 13]. One kind of partial exchangeability, known as 'Markov exchangeability', allows outcomes to depend on previous trials [14, 16, 17, 25, 30, 35]. The type of partial exchangeability most relevant to our analogical inductive logic was investigated by de Finetti himself [10, 11]. Consider a situation where outcomes can be of different types; e.g., coin flips with two coins, or medical trials with men and women. Then one may not be willing to judge outcomes to be exchangeable across types but only within types. There is a representation theorem for this kind of partial exchangeability, from which predictive conditional probabilities can be derived [10]. The representation is very similar to (3). Probabilities are again mixtures of independent trials, but now trials need not be identically distributed; they are identically distributed within types, but need not be so be across types.

de Finetti viewed partial exchangeability as a type of analogical inference. Take the example of flipping two coins. The coins are judged to be similar but not indistinguishable from each other. Because of the analogy between the two coins, observations from one coin should have some influence on predictions for the other coin. The analogy comes from particular prior distributions on the chances in the mixture of the representation theorem. The biases of the two coins may be chosen dependently, but then trials are independent. Thus this kind of analogy influence does not persist for very long. This is also a feature of some analogical inductive logics considered in the next section.

4. THE PROBLEM OF ANALOGICAL INDUCTIVE INFERENCE

Carnap's basic system of inductive logic can express analogical influences only to a limited degree [6, 28, 29]. There have been many attempts to extend Carnap's original system, and the literature on analogical inductive logic includes many valuable contributions [1, 5, 8, 15, 18, 19, 23, 26, 27, 12, 28, 29, 31, 32]. I am going to discuss some of those contributions in order to motivate my own.

The biggest obstacle to analogical inference in Carnap's system is Johnson's sufficientness postulate (1). Johnson's sufficientness postulate makes it impossible that counts n_k of outcomes k other than i influence the predictive probability of i. Skyrms [31] suggests an extension of Carnapian inductive logic that keeps exchangeability but drops Johnson's sufficientness postulate. Skyrms' proposal is further studied and extended in [15] and [19], and a similar model is developed for a different context (two families of predicates) in [29]. The basic idea is to use mixtures of inductive methods (2) in order to account for initial analogies between outcomes. This is equivalent to considering mixtures of Dirichlet distributions instead of Dirichlet distributions in the de Finetti representation. Skyrms discusses this idea in terms of a wheel of fortune, where observations of an outcome should also increase the predictive probability of nearby outcomes. Using an appropriate mixture of Dirichlet priors makes this possible. The resulting probability distributions are exchangeable but violate Johnson's sufficientness postulate.

The analogy influence exhibited by these kinds of inductive systems is transient. This is due to the fact that the corresponding prior probabilities are exchangeable. Exchangeability implies that the counts of one outcome can only have an indirect effect on the predictive probabilities of other outcomes. To see this, suppose that an outcome k is followed by an outcome i. Then exchanging i with some arbitrary outcome in the past does not affect the joint probability. Thus, the effect of counts of k outcomes affects the probability of i outcomes indirectly via the initial parameters in the mixture of Carnapian inductive logics.

In order to get systems that exhibit a more permanent analogy influence, exchangeability has to be dropped in addition to Johnson's sufficientness postulate. The inductive systems of Costantini [8], Kuipers [23], Niiniluoto [28] and, to a certain extent, Spohn [32] develop inductive logics of this type. In these models, the predictive probabilities for outcome i do not just contain the counts n_i but may also have terms with counts n_k of other outcomes. Each of these systems is interesting in its own right, but for none of them is it clear what the underlying symmetry assumptions are, or whether they exhibit interesting symmetries at all, and thus they seem a bit ad hoc.

Another criticism of some of these inductive methods was put forward by Spohn [32] and is also expressed by Costantini [8]. Because counts of all outcomes may explicitly influence the predictive probabilities of an outcome i, the corresponding inductive logics generally violate a postulate known as 'Reichenbach's axiom'. Reichenbach's axiom says that predictive probabilities have to converge to limiting relative frequencies of sample outcomes, provided that the limit exists. That is, if $X_1, X_2, \ldots$ is an infinite sequence of outcomes such that the limit n_i/n exists as $n \to \infty$, then

$$\lim_{n\to\infty} \mathbb{P}[X_{n+1} = i | X_1, \ldots, X_n] = p.$$

Besides Spohn's own system, Carnap's basic system and Skyrms' analogical system meet Reichenbach's axiom.

I think this critique misses the point of certain forms of inductive inference. The inductive logics of Costantini and Niiniluoto may be appropriate when there are underlying probabilistic dependencies between the outcomes. If these dependencies are persistent, then Reichenbach's axiom should not hold. The dependencies will not be reflected in relative frequencies of outcomes, while predictive probabilities should make use of known dependencies. I discuss this point further in the context of our analogical inductive logic.

It is not known whether the inductive methods discussed so far can be derived from a set of axioms analogous to those underlying the Johnson-Carnap system. This is a significant gap in our knowledge. The set of axioms from which the Johnson-Carnap continuum of inductive methods (2) is derivable completely specifies inductive assumptions at the observational level, making it easy to determine whether one's priors conform to them. None of the above models of inductive inference has been treated within this Carnapian paradigm. Maher's inductive logic is something of an exception [26, 27]. He presents a set of axioms for an inductive logic with two families of predicates. Maher himself discusses problems for the extension of the inductive logic to predicate families containing more than two predicates [27], so I confine my attention to the case of two predicates. I think that already in this case one point is in need of clarification.

Here is a brief overview of Maher's proposal. Suppose we have two families with two predicates. In the language of random variables, this means that we have two sequences of random variables $V_1, V_2, \ldots$ and $W_1, W_2, \ldots$, where each random variable can take on two different values (the possible values being different for the V's and the W's). For instance, the first sequence might record whether the coin lands heads and tails, and the second sequence may state whether the coin is flipped with the right or the left hand. Maher then considers the so-called 'Q-predicates' ('state descriptions' in Carnap's terminology). The Q-predicates are all possible combinations of basic predicates from the two families. Again in the language of random variables, this means that we consider the sequence of pairs $Z_n = (V_n, W_n)$. The random vector Z_n takes on pairs of values. Since the random variables V_n and W_n are binary, Z_n can take on four values.

Maher [26] assumes that the infinite sequence $Z_1, Z_2, \ldots$ is exchangeable. It follows from this that its probability distribution has a de Finetti representation. Maher's basic idea can then be described as follows. The de Finetti representation implies that we can construct the probability distribution of $Z_1, Z_2, \ldots$ by putting a prior distribution over the set of possible chances. Since the Z_n can take on four different values, the set of possible chances is the three-dimensional simplex $\Delta_4 = \{(x_1, \ldots, x_4) \in \mathbb{R}^4 | x_1, \ldots, x_4 \geq 0, x_1 + \ldots + x_4 = 1\}$. Following an idea by Carnap [6], Maher considers the subset of probability distributions in Δ_4 where the two families of predicates are probabilistically independent. This is the set of all $(x_1, \ldots, x_4) \in \Delta_4$ such that $x_1 = (x_1 + x_2)(x_1 + x_3)$, which defines a two-dimensional surface in Δ_4 that is known as the 'Wright manifold' in population genetics.[4]

If the prior on Δ_4 is a Dirichlet distribution, as in Carnap's basic system, then any two-dimensional surface in Δ_4 has probability zero, since the Dirichlet distribution is absolutely continuous with respect to Lebesgue measure on Δ_4. Thus, the Wright manifold has probability zero. Now, Carnap and Maher propose to look at a mixture between a Dirichlet distribution and a distribution that puts full weight on the Wright manifold. The resulting inductive logic is a mixture of Carnap's basic system on the random variables Z_n and the product of Carnap's basic systems on the random variables V_n and W_n. The former terms correspond to the hypothesis that the two predicate families are dependent and the product of the latter two terms to the hypothesis that they are independent. Using the de Finetti representation, Maher also provides an axiomatic basis from which this inductive method can be derived. He also shows with the help of examples that the resulting system seems to lead to plausible numerical results that capture certain analogy influences.

What type of analogy influences is this model supposed to capture? Maher wants to say that some of the Q-predicates are more similar than others, namely those that share at least one underlying predicate from the two families. If we denote the four combinations of values by $Q_1 = (0, 0)$, $Q_2 = (1, 0)$, $Q_3 = (0, 1)$ and $Q_4 = (1, 1)$ then Q_1 is similar to Q_2 and Q_3, Q_2 to Q_1 and Q_4, Q_3 is similar to Q_1 and Q_4, and Q_4 is similar to Q_2 and Q_3. Maher's goal is to have an inductive logic that respects the analogies based on these similarities. But it is difficult to see the reason why placing positive prior probability on the Wright manifold should achieve this. There is no straightforward relationship between considering the two predicate families as independent and the intended analogies.

The one reason I can see is the following. The similarity relationships between the Q-predicates described in the previous paragraph yield four edges in Δ_4 between the vertices that are considered similar. These edges are part of the Wrigth manifold. If one wishes to reflect the analogies between the Q-predicates in one's prior, then one's prior distribution over Δ_4 should, presumably, place a sufficient amount of probability weight close to the four edges. One way to achieve this is by distributing probabilities in an appropriate way on the Wright manifold. But this is neither necessary nor sufficient. We can endow the Wright manifold by assigning positive probability only to the barycenter of Δ_4 (which is an element of the Wright manifold) and probability zero to all the other points in the Wright manifold. In this case, the overall prior over Δ_4 may not place the required probability weight close to the four edges. On the other hand, we may do exactly that without having to assign positive probability to the Wright manifold. Thus, even though it may work in some cases, assigning positive probability to the Wright manifold does not seem to be a principled solution to the analogy problem, which would characterize priors over Δ_4 that assign a sufficient probability weight to the four edges between analogous Q-predicates.

5. EXTENDING PARTIAL EXCHANGEABILITY

The brief discussion in the previous section should make it clear that there are many forms of analogical inference. Each form of analogical inference merits study, and existing inductive logics vary in their degree of solving analogical inference problems successfully. In the remainder of this paper I would like to propose one form of analogical inductive inference that is based on de Finetti's ideas about partial exchangeability and that can be solved within the Carnapian paradigm.

Recall that partial exchangeability looks at situations with outcomes of different types. This inductive situation can be illustrated with an example that Achinstein used to criticize Carnap's original inductive logic [1]. In this example we observe whether or not different types of metal conduct electricity. We might, for instance, look at osmium, platinum and rhodium. These three metals are the types in de Finetti's setup. Each type may or may not conduct electricity. This defines two outcomes. The analogy between types comes from the fact that they share certain significant chemical properties. Because of the analogy between types, it is reasonable to think that instances where osmium and rhodium where observed to conduct electricity are relevant for predictions of whether platinum conducts electricity. In

[4] After the population geneticist Sewall Wright. The Wright manifold is the set of probabilities that make the alleles at different genetic loci independent.

this case, de Finetti's theory of partial exchangeability may be applied with a prior that reflects these analogies.

Partial exchangeability has a similar effect on analogical inferences as exchangeability: analogy is transient and vanishes in the limit. This makes sense in the example of flipping two coins. The similarity between the two coins may influence one's early judgements, but if there are no underlying dependencies between the coins the influence of similarity judgements will diminish. This is reflected by the fact that Reichenbach's axiom holds for predictive probabilities. But what if there are persistent dependencies between types? This might arguably be the case in the example of whether different metals conduct electricity, since there presumably is an underlying common cause for the relevant outcome. Another example can be constructed by considering the success of medical trials among males and females. The types are male and female, and the outcomes (in the simplest case) are whether the trial was successful or not. Now, there might be an underlying chancy dependency between types that is influenced by environmental and other factors. If this dependency is permanent, this should be reflected in the analogical inductive logic.

How might such an inductive logic look like? The basic setup has a sequence of outcomes $X_1, X_2, \ldots$ and a sequence of types $Y_1, Y_2, \ldots$. Suppose, for simplicity, that there are only two types. Predictive probabilities concern future outcomes and not future types. The predictive probability of observing outcome i given that it is of type 1 and given past observations may be given by

$$\mathbb{P}[X_{N+1} = i | \mathbf{X}_N, \mathbf{Y}_N, Y_{N+1} = 1] = \frac{n_{i1} + \beta n_{i2} + \alpha_{i1}}{N_1 + \beta N_2 + \sum_j \alpha_{j1}}. \tag{4}$$

In this formula, $\mathbf{X}_n = (X_1, \ldots X_N), \mathbf{Y}_N = (Y_1, \ldots, Y_N)$ are the past observations of outcomes and types; n_{ij} is the number of outcomes i of types j; and N_1 and N_2 are the total number of observations of type 1 and 2. The α parameters have the same meaning as in Carnap's basic system (2). The β parameter expresses the analogy influence of observations of type 2 on observations of type 1. If β is positive, then i observations of type 2 will have a positive influence on the predictive probability. This indicates a judgement of positive analogy between types. Moreover, analogy is permanent—since β is a constant, the analogy influence of type 1 on type 2 does not vanish as n increases.

There are many ways in which the qualitative features of the predictive probability (4) could be formalized. Is (4) just a formula that exhibits some resemblance to Carnap's original system? Or is there some underlying rationale? To see what is going on, notice, in the first place, that de Finetti's notion of partial exchangeability will not in general allow predictive probabilities to be of the form as given in (4). Partial exchangeability implies the following. Suppose that $X_{N+1} = i, X_{N+2} = k, X_{N+3} = j$. The predictive probability of this sequence of outcomes, given the past and the sequence of types $Y_{N+1} = 1, Y_{N+2} = 2, Y_{N+3} = 1$, is equal to the predictive probability of the sequence $X_{N+1} = j, X_{N+2} = k, X_{N+3} = i$ (in order to get from the first sequence of outcomes to the second we only exchange two outcomes within the same type). Now suppose that $k = j$. Then the first sequence of outcomes is $X_{N+1} = i, X_{N+2} = j, X_{N+3} = j$ and the second is $X_{N+1} = j, X_{N+2} = j, X_{N+3} = i$. It is difficult to see how in this case counts of outcome j of type 2 can have a constant influence on the predictive probability of outcomes j of type 1. If it had, its effect would have to be balanced exactly against the joint probability for the second sequence, which may not work in general.[5]

The same issue does not arise if $k \neq i, j$. Thus, it seems reasonable to weaken partial exchangeability in order to allow for persistent analogical influences. We let $p^n_{ikj,st} = \mathbb{P}[X_{N+1} = i, X_{N+2} = k, X_{N+3} = j | \mathbf{X}_n, \mathbf{Y}_n, Y_{N+1} = s, Y_{N+2} = t, Y_{N+3} = s]$. Then generalized partial exchangeability requires, in the first place, that

$$p^n_{ikj,st} = p^n_{jki,st}$$

whenever $k \neq i, j$ (if $k = i$ or $k = j$, equality may but need not hold). Furthermore, let $p^n_{ij,s} = \mathbb{P}[X_{N+1} = i, X_{N+2} = j | \mathbf{X}_n, \mathbf{Y}_n, Y_{N+1} = s, Y_{N+2} = s]$. Then generalized partial exchangeability requires, in the second place, that

$$p^n_{ij,s} = p^n_{ji,s}$$

The next section is devoted to showing that generalized partial exchangeability, together with some further assumptions, leads to an interesting analogical inductive logic.

6. A NEW ANALOGICAL INDUCTIVE LOGIC

The most important additional assumption that we need is a modification of Johnson's sufficientness postulate:

$$\mathbb{P}[X_{N+1} = i | \mathbf{X}_N, \mathbf{Y}_N, Y_{N+1} = j] = f_{ij}(n_{i1}, n_{i2}, N_1, N_2) \tag{5}$$

For simplicity, we continue assuming that there are only two types (for a generalization to a finite number of types, see [20]). The modified sufficientness postulate says that predictive probabilities for an outcome i depend on i, its type, as well as on the observed counts of i outcomes of both types. This is a natural way to allow for analogical influences between types.

We also need two technical postulates. The first one is a regularity assumption to the effect that all finite sequences of types and outcomes have positive probability; i.e., every finite pair of sequences $X_1, \ldots, X_N, Y_1, \ldots, Y_N$ has positive probability. Finally, we assume that future types do not give information about the outcome of the next trial. More specifically,

$$\begin{aligned} &\mathbb{P}[X_{N+1} = i | X_1, \ldots, X_N, Y_{N+1} = j] \\ &\quad = \mathbb{P}[X_{N+1} = i | X_1, \ldots, X_N, Y_{N+1} = j, Y_{N+2} = k] \\ &\quad = \mathbb{P}[X_{N+1} = i | X_1, \ldots, X_N, Y_{N+1} = j, Y_{N+2} = k, Y_{N+3} = l]. \end{aligned} \tag{6}$$

This condition is a significant restriction for the applicability of our inductive logic. For example, think of types as different medical treatments (as in a bandit problem) and of outcomes as success or failure. Then a success on the next trial might not be probabilistically independent of future treatments.

Suppose now that $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are two infinite sequences of outcomes and types for which the foregoing assumptions hold (generalized partial exchangeability, modified sufficientness postulate, regularity, and conditional independence (6)). Suppose, in addition, that trials within types are not independent, and that there are at least three outcomes.[6] Then the following theorem is true:

[5] For a precise statement, see my [20], especially Corollary 2.

[6] Assuming independence has the same reason as in the case

Theorem 1. *There exist positive constants α_{ij} and non-negative constants β, γ such that $N_1 + \beta N_2 + \sum_i \alpha_{i1} \neq 0, N_2 + \gamma N_1 + \sum_i \alpha_{i2} \neq 0$ and*

$$\mathbb{P}[X_{N+1} = i | \mathbf{X}_N, \mathbf{Y}_N, Y_{N+1} = 1] = \frac{n_{i1} + \beta n_{i2} + \alpha_{i1}}{N_1 + \beta N_2 + \sum_i \alpha_{i1}}$$

$$\mathbb{P}[X_{N+1} = i | \mathbf{X}_N, \mathbf{Y}_N, Y_{N+1} = 2] = \frac{n_{i2} + \gamma n_{i1} + \alpha_{i2}}{N_2 + \gamma N_1 + \sum_i \alpha_{i2}}$$

for all N and all $0 \leq n_{ij} \leq N_j$.

This theorem follows from a more general result in my [20] where I prove these assertions for more than two types and allow the total number of trials to be finite.

The sequence of predictive probabilities can be generated by an urn model (just like the predictive probabilities of Carnap's basic system are generated by a Polya urn). Since the predictive probabilities of our new inductive logic do not fix the probabilities of types, we may first choose a sequence of types at random from a distribution that assigns positive probability to each finite sequence of types. Assume that we also have an urn for each type containing balls labelled by the outcomes. The initial distribution of balls in urn j depends on the prior parameters α_{ij}. We now start choosing balls from urns following the sequence of types. Whenever we choose a ball from an urn, we put it back together with another label. If the urn is of type 1, we put a ball with weight β into the urn associated with type 2.

The most important difference between our new inductive logic and Carnap's basic system (2) are the parameters β, γ. Are there any good reasons to think that they are analogy parameters? Let me mention two. First, it can be shown that β is positive if

$$P[X_2 = i | X_1 = i, Y_1 = 2, Y_2 = 1] > \mathbb{P}[X_1 = i | Y_1 = 1].$$

Furthermore, β increases as $P[X_2 = i | X_1 = i, Y_1 = 2, Y_2 = 1]$ approaches 1.[7] This means that we have analogy effects of type 2 on type 1 if observing an outcome of type 2 makes it sufficiently more likely to observe the same outcome of type 1. This is what one would expect of an analogical inference.

The second reason becomes relevant if there are more than two types. Consider the analogy parameters of two types with respect to a third one. If one parameter is larger than the other, then observing outcomes of the former type raises the probability of outcomes of the third type more than observing outcomes of the second type.[8]

The inductive logic of Theorem 1 is open to various interpretations. If we interpret the parameters β and γ as analogy parameters, then it is plausible to require that $\beta, \gamma \leq 1$ since, arguably, every type is maximally analogous to itself. This idea can be captured by another postulate:

$$\begin{aligned}\mathbb{P}[X_2 = i | X_1 = i, Y_1 = j, Y_2 = j] \\ \geq \mathbb{P}[X_2 = i | X_1 = i, Y_1 = k, Y_2 = j]\end{aligned}$$

This says that an observation of an outcome i of type j never has a lower effect on the predictive probability of that outcome when it is of type j than observing an outcome i of another type. It is easy to see that this forces the analogy parameters β, γ to be between zero and one.

But we may also think of types in terms of different information sources that are used to predict probabilities of outcomes. In this case, β and γ express judgements about the reliability of the two sources. Consequently, if $\beta > 1$ the agent believes that the second information source is more trustworthy than the first one and that, accordingly, information from type 2 observations should have more weight.

One feature of the inductive logic of Theorem 1 was already discussed earlier in a different context. Our new inductive logic violates Reichenbach's axiom whenever the analogy parameters β and γ are positive. In this case, predictive probabilities converge to a convex combination of relative frequencies of outcomes of the two different types. As remarked earlier, if the underlying process is not assumed to be essentially independent, this is what one should expect. Our inductive logic allows types to be probabilistically dependent throughout the process of observation, and so observations from other types don't necessarily cease to be relevant for predictive probabilities of one particular type. Thus, Reichenbach's axiom should not be postulated for this case.

7. CONCLUSION

One of the biggest advantages of our inductive logic is that there is a precise set of conditions from which it can be derived. These conditions can be thought of as the inductive assumptions that make the use of our analogical inductive logic adequate, provided that they are thought to be true. For most other analogical inductive logics the underlying assumptions are not as clear, which makes it difficult to apply them.

What I wish to emphasize is that there are different ways to reason analogically. Accordingly, there is going to be a variety of legitimate analogical inductive logics, and not just the one inductive logic that fully captures analogical reasoning. One basic distinguishing feature is suggested by the foregoing discussion. There are, on the one hand, inductive logics where analogies reflect initial similarities but are washed out with increasing information. On the other hand, there are permanent analogical inferences such as in our inductive logic. Here, analogy persists with increasing information. Which type of analogy is appropriate depends on one's inductive assumptions.

8. REFERENCES

[1] P. Achinstein. Variety and analogy in confirmation theory. *Philosophy of Science*, 30:207–221, 1963.

[2] R. Carnap. *Logical Foundations of Probability.* University of Chicago Press, Chicago, 1950.

[3] R. Carnap. *The Continuum of Inductive Methods.* University of Chicago Press, Chicago, 1952.

[4] R. Carnap. A basic system of inductive logic, part 1. In Rudolf Carnap and Richard C. Jeffrey, editors, *Studies in Inductive Logic and Probability I*, pages 33–165. University of California Press, Los Angeles, 1971.

[5] R. Carnap. A basic system of inductive logic, part 2. In R. C. Jeffrey, editor, *Studies in Inductive Logic and*

of the Johnson-Carnap continuum—independence means that there is no inductive learning. Since the sufficientness postulate is empty if there are only two outcomes, this case has to be treated separately, for example by assuming additivity of predictive probabilities. An alternative approach is proposed in [7].

[7] Similar relations hold for γ; see [20].

[8] See Proposition 1 in [20].

Probability II, pages 7–155. University of California Press, Los Angeles, 1980.

[6] R. Carnap and W. Stegmüller. *Induktive Logik und Wahrscheinlichkeit*. Springer, Wien, 1959.

[7] D. Costantini. The relevance quotient. *Erkenntnis*, 14:149–157, 1979.

[8] D. Costantini. Analogy by similarity. *Erkenntnis*, 20:103–114, 1983.

[9] B. de Finetti. La prevision: ses lois logiques ses sources subjectives. *Annales d l'institut Henri Poincaré*, 7:1–68, 1937. Translated in Kyburg, H. E. and Smokler, H. E., editors, *Studies in Subjective Probability*, pages 93–158, Wiley, New York, 1964.

[10] B. de Finetti. Sur la condition d'equivalence partielle. In *Actualités Scientifiques et Industrielles No. 739: Colloques consacré à la théorie des probabilités, VIième partie*, pages 5–18. Paris, 1938. Translated in Jeffrey, R. C., editor, *Studies in Inductive Logic and Probability II*, pages 193–205, University of California Press, Los Angeles, 1980.

[11] B. de Finetti. La probabilita e la statistica nei raporti con l'induzione, secondo i dwersi punti di vista. In *Corso C.I.M.E su Induzione e Statistica*. Cremones, Rome, 1959. Translated in de Finetti, B, *Probability, Induction and Statistics*, chapter 9, Wiley, New York, 1974.

[12] M. C. di Maio. Predictive probability and analogy by similarity in inductive logic. *Erkenntnis*, 43:369–394, 1995.

[13] P. Diaconis and D. Freedman. De Finetti's generalizations of exchangeability. In R. C. Jeffrey, editor, *Studies in Inductive Logic and Probability II*, pages 233–249. University of California Press, Los Angeles, 1980.

[14] P. Diaconis and D. Freedman. De Finetti's theorem for Markov chains. *Annals of Probability*, 8:115–130, 1980.

[15] R. Festa. Analogy and exchangeability in predictive inferences. *Erkenntnis*, 45:89–112, 1997.

[16] S. Fortini, L. Ladelli, G. Petris, and E. Regazzini. On mixtures of distributions of Markov chains. *Stochastic Processes and their Applications*, 100:147–165, 2002.

[17] D. Freedman. Mixtures of Markov processes. *Annals of Mathematical Statistics*, 33:114–118, 1962.

[18] M. Hesse. Analogy and confirmation theory. *Philosophy of Science*, 31:319–324, 1964.

[19] A. Hill and J. Paris. An analogy principle in inductive logic. *Annals of Pure and Applied Logic*, 64:1293–1321, 2013.

[20] S. M. Huttegger. Analogical predictive probabilities. Manuscript, University of California at Irvine, 2015.

[21] W. E. Johnson. *Logic, Part III: The Logical Foundations of Science*. Cambridge University Press, Cambridge, UK, 1924.

[22] W. E. Johnson. Probability: The deductive and inductive problems. *Mind*, 41:409–423, 1932.

[23] T. Kuipers. Two types of inductive analogy by similarity. *Erkenntnis*, 21:63–87, 1984.

[24] T. A. F. Kuipers. *Studies in Inductive Probability and Rational Expectation*. D. Reidel, Dordrecht, 1978.

[25] T. A. F. Kuipers. Inductive analogy by similarity and proximity. In D. H. Helman, editor, *Analogical Reasoning*. Kluwer, Dordrecht, 1988.

[26] P. Maher. Probabilities for two properties. *Erkenntnis*, 52:63–91, 2000.

[27] P. Maher. Probabilities for multiple properties: The models of Hesse and Carnap and Kemeny. *Erkenntnis*, 55:183–216, 2001.

[28] I. Niiniluoto. Analogy and inductive logic. *Erkenntnis*, 16:1–34, 1981.

[29] J.-W. Romeijn. Analogical predictions for explicit similarity. *Erkenntnis*, 2006:253–280, 2006.

[30] B. Skyrms. Inductive logic for Markov chains. *Erkenntnis*, 35:439–460, 1991.

[31] B. Skyrms. Analogy by similarity in hyper-Carnapian inductive logic. In J. Earman, A. I. Janis, G. Massey, and N. Rescher, editors, *Philosophical Problems of the Internal and External Worlds*, pages 273–283. University of Pittsburgh Press, Pittsburgh, 1993.

[32] W. Spohn. Analogy and inductive logic: A note on Niiniluoto. *Erkenntnis*, 16:35–52, 1981.

[33] S. L. Zabell. W. E. Johnson's "sufficientness" postulate. *The Annals of Statistics*, 10:1091–1099, 1982.

[34] S. L. Zabell. The Rule of Succession. *Erkenntnis*, 31:283–321, 1989.

[35] S. L. Zabell. Characterizing Markov exchangeable sequences. *Journal of Theoretical Probability*, 8:175–178, 1995.

[36] S. L. Zabell. Carnap and the logic of inductive inference. In *Handbook of the History of Logic*, pages 265–309. Elsevier, Amsterdam, 2011.

Human-Agent Decision-making: Combining Theory and Practice

[Extended Abstract]

Sarit Kraus
Dept. of Computer Science
Bar-Ilan University
Ramat-Gan, Israel
sarit@cs.biu.ac.il

ABSTRACT

Extensive work has been conducted both in game theory and logic to model strategic interaction. An important question is whether we can use these theories to design agents for interacting with people? On the one hand, they provide a formal design specification for agent strategies. On the other hand, people do not necessarily adhere to playing in accordance with these strategies, and their behavior is affected by a multitude of social and psychological factors. In this paper we will consider the question of whether strategies implied by theories of strategic behavior can be used by automated agents that interact proficiently with people. We will focus on automated agents that we built that need to interact with people in two negotiation settings: bargaining and deliberation. For bargaining we will study game-theory based equilibrium agents and for argumentation we will discuss logic-based argumentation theory. We will also consider security games and persuasion games and will discuss the benefits of using equilibrium based agents.

Categories and Subject Descriptors

I.2 [**ARTIFICIAL INTELLIGENCE**]: Miscellaneous

Keywords

Intelligent Agents

1. INTRODUCTION

Agents that interact proficiently with people may be useful for training [27], supporting [23, 12, 11, 1] and even replacing people in many applications [10, 22].

We are considering the agent-human interactions as being a *strategic* activity [14]. That is, we assume that when the automated agent engages in the interaction, it should act as best it can to realize its preferences. *Game theory* is the mathematical theory of strategic decision-making [28] and thus it seems that game theory might be an appropriate analytical tool for understanding how a strategic agent can and should act, and might also be useful in both the design of automated agents and protocols for the interactions. However, game theory assumes that all players will act as best they can to realize their preferences. Unfortunately, humans tend to make mistakes, and they are affected by cognitive, social and cultural factors [8, 25, 4]. In particular, people's observed behavior does not correspond to game theory-based equilibrium strategies [13, 29].

Another approach for the development of automated agents is the (non-classical)-logic approach. The agent is given a logical representation of its environment and its desired goals, and it reasons logically in order to generate its activities. When interacting with people, the environment consists also of the human model. Yet, modeling people's behavior is a big challenge. We have incomplete information about the person's preferences, and we have to cope with the uncertainties inherent in human decision-making and behavior. Human behavior is diverse, and cannot be satisfactorily captured by a simple abstract model. In particular, human decision-making tends to be very noisy: a person may make different strategic decisions in similar situations.

In this paper we survey briefly a few of the agents that we built over the years that interact proficiently with people. In most of the cases, deploying only a game-theory approach or logical-based approach was not beneficial. Heuristics and machine-learning techniques were augmented into the formal models to lead to agents that interact proficiently with people. We will discuss three negotiation settings: multi-issue negotiations, games where the players interleave negotiations with resource exchange while attempting to satisfy their goals and argumentation settings. Finally, we will discuss security games.

2. MULTI-ISSUE NEGOTIATIONS

Over the years we designed and implemented several automated agents for multi-issue negotiations. In multi-issue negotiations the players need to reach an agreement on several issues. Each issue is associated with a set of possible values and the players need to agree on a specific value for each issue. The negotiations can end with the negotiators signing an agreement or with one of the sides opting out of the negotiations. In addition, if the crisis does not end within a pre-specified deadline then the status quo is implemented. Each outcome of the negotiations is associated with a utility score for both players. A summary of our agents is presented in Table 1.

2.1 EQH agent

The first agent, EQH, that we developed was for crisis scenarios and the setting was quite complex [24]. In addition to the message exchange in a semi-structured language, players could take actions during the negotiations and agreements were not enforceable. In particular, opting out in a crisis is a stochastic action and thus the agents are uncertain about

Table 1: Multi-issue negotiations

Settings	Agent Name	Agent Properties	Scenarios	Significance vs people
Bilateral, single-issue, full information, complex actions, agreements not enforceable	EQH	SPE with manually designed heuristics	fishing dispute	One role
Bilateral, uncertainty, multi-issue	QO-agent	Qualitative decision-making Non-deterministic behavior	job interview tobacco	One role
	KBAgent	Machine learning, qualitative decision-making non-deterministic behavior	job interview tobacco	Both roles
	NegoChat	KBagent algorithms, AAT Anchoring, NLP module	job interview	Both roles

the result. In addition to the main issue of the negotiation or opting out, there are various other parameters of an agent's action. These parameters influence the utility of the negotiators from the crisis. Time plays an important role in the crisis [42]. The specific scenario we used for the experimental study was a fishing dispute between Canada and Spain. We formalized the crisis scenario as a game and identified a subgame-perfect equilibrium. We ran preliminary experiments when the automated agent followed its subgame perfect equilibrium strategy. However, the human negotiators who negotiated with it became frustrated and the negotiation often ended with no agreement. The frustration of the human negotiators was mainly due the lack of flexibility of the agent. Since the proposed and accepted agreements of the subgame perfect negotiation did not change over the negotiation time, the agent did not compromise.

To address this limitation of the equilibrium-based agent, we incorporated several heuristics to the EQH agent. We allowed the owner of the agent to determine the way the agent will deviate from the equilibrium strategies by determining parameters that influence the agent's behavior which are instantiated before the beginning of negotiations. In order to provide the agent with some flexibility when playing against people, we allowed the agent to consent to agreements that have a lower utility than it would have obtained according to the relevant subgame perfect equilibrium strategy agreement. Therefore we added the margin parameter that determines the largest number of points lower than the desired utility value to which the agent will agree.

An additional parameter is the number of negotiation units by which the agent will increase or decrease its first offer from the agreement specified in its equilibrium strategy. Human negotiators usually begin negotiations with an offer higher (or lower, depending on the negotiator's role) than the value they would eventually like to reach at the end of negotiations. This leaves bargaining space and our agent uses this type of strategy. Another parameter indicates whether the agent will send the first message in the negotiation or will wait for its opponent to make the first offer. The default value of this parameter, following some literature recommendations [16], was that the agent will send the first offer, since we wanted a trigger to initiate negotiations with the other agent.

Another heuristic concerns opting out. Given our assumptions, while rational agents will not opt out, people may opt out. If the agent's expected utility from opting out is higher than its expected utility from its opponent opting out, it will try to predict whether its opponent is going to opt out. If so, it will opt out first. The heuristic for the prediction of whether an opponent will opt out is based on the messages sent by the opponent. For example, when a threatening message is received, or when a comment message indicating that the negotiations are heading in a dangerous direction is received, the estimation that the opponent may opt out increases.

We ran extensive experiments for evaluating the equilibrium agent with the heuristics (EQH agent) [24]. We compared the results of the humans to those of the agents and concluded that the EQH agent received a higher utility score playing both roles, but the results were only statistically significant when the agent played just one of the roles. Furthermore, when an agent participates in a negotiation, the sum of the utilities are significantly higher than when two humans play since the agent always proposes Pareto-optimal offers while people reach agreements that are not.

While the EQH agent was based on the subgame perfect equilibrium strategies, it required the introduction of many heuristics, and its success compared with people was only in one role. The main open question is whether it is possible to provide formal methodology that will lead to an agent that is similar to the EQH without the need to manually design the EQH heuristics. Furthermore, we are aiming for an agent that can achieve a significantly higher utility score than people in both roles. Toward this challenges, we next tried to use a qualtative approach, to introduce incomplete information into the environment and to improve the agent's results in both roles.

2.2 QOagent and KBagent

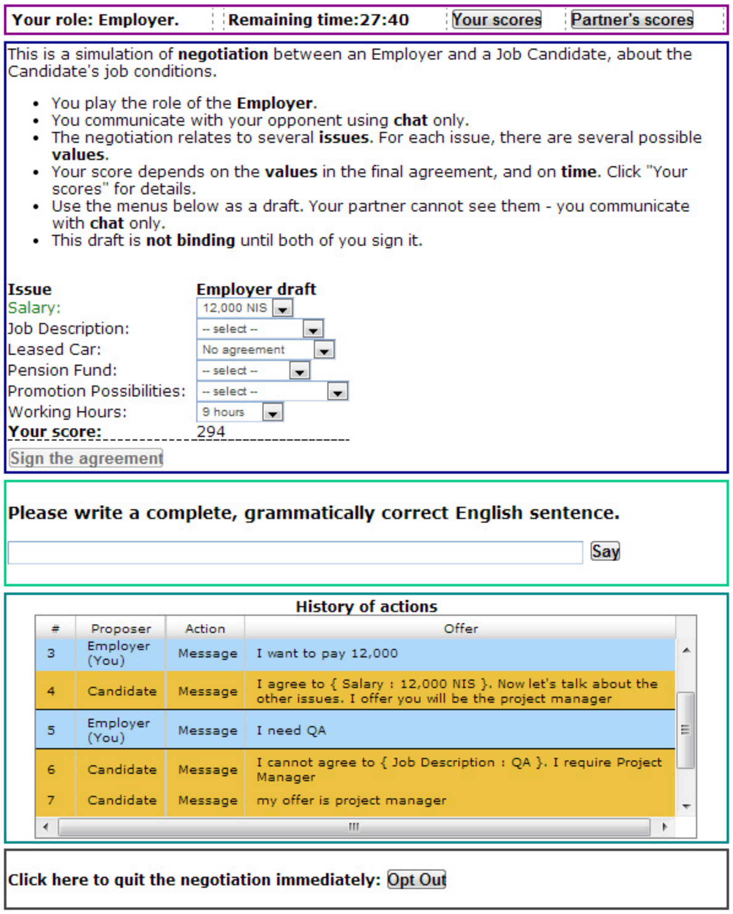
The QOagent was designed to interact with people in environments of bilateral negotiations with incomplete information when the agreements consist of multiple issues [26]. With respect to incomplete information, each negotiator keeps his preferences private, though the preferences might be inferred from the actions of each side (e.g., offers made or responses to offers proposed). Incomplete information is expressed as uncertainty regarding the utility preferences of the opponent, and it is assumed that there is a finite set of different negotiator types. These types are associated with different additive utility functions (e.g., one type might have a long term orientation regarding the final agreement, while the other type might have a more constrained orientation). Lastly, the negotiation is conducted once with each opponent. The experiments were run on two distinct domains. In the first domain, England and Zimbabwe negotiate in order to reach an agreement evolving from the World Health Organization's Framework Convention on Tobacco Control, the world's first public health treaty. In the second domain a negotiation takes place after a successful job interview between an employer and a job candidate.

We first formalized the scenario as a Bayesian game and computed the Bayesian Nash equilibrium. Though we did not run simulations of the Bayesian Nash equilibrium agent against human negotiators, we ran two humans negotiations. We found out that the opponent's utility score from the offers suggested by the equilibrium agent are much lower than the final utility values of the human negotiations. By also analyzing the simulation process of the human negotiations, we deduced that without incorporating any heuristics into the equilibrium agent, the human players would not have accepted the offers proposed by it which will lead to low utility scores for the equilibrium agent, similar to the low score of the equilibrium agent in the fishing dispute.

Therefore, we developed the QOAgent. For the decision-making process, the approach used by the QOAgent tries to take the utility of both sides into consideration. While the QOAgent's model applies utility functions, it is based on a non-classical decision-making method, rather than focusing on maximizing the expected utility: the maximin function and a qualitative valuation of offers. Using these methods, the QOAgent generates offers and decides whether to accept or reject proposals it has received. As for incomplete information, the QOAgent tackles this problem using a simple Bayesian update mechanism. After each action, this mechanism tries to infer which negotiator type best suits the opponent.

The effectiveness of this method was demonstrated through extensive empirical experiments by [26].

The results of the experiments showed that the automated agent achieved higher utility scores than the human counterpart. This can be explained by the nature of our agent both in reference to accepting offers and generating offers. Using the decision-making mechanism we allow the agent to propose agreements that are good for it, but also reasonable for its opponent. In addition, the automated agent makes straightforward calculations. It evaluates the offer based on its attributes, and not based on its content. In addition, it also places more weight on the fact that it loses or gains as time advances. This is not the case, however, when analyzing the logs of the people. It seems that people put more weight on the content of the offer than on its value. This was more evident in the Job Candidate domain with which the human subjects could more easily identify. Yet, this does not explain why, in both domains, similar to the EQH agent experiments, these results are significant only for one of the sides. In the England-Zimbabwe domain, the results are significant when the agent played the role of England, while in the Job Candidate domain these results are significant when it played the role of the job candidate.



Your role: Employer. Remaining time:27:40

This is a simulation of **negotiation** between an Employer and a Job Candidate, about the Candidate's job conditions.

- You play the role of the **Employer**.
- You communicate with your opponent using **chat** only.
- The negotiation relates to several **issues**. For each issue, there are several possible **values**.
- Your score depends on the **values** in the final agreement, and on **time**. Click "Your scores" for details.
- Use the menus below as a draft. Your partner cannot see them - you communicate with **chat** only.
- This draft is **not binding** until both of you sign it.

Issue	**Employer draft**
Salary:	12,000 NIS
Job Description:	-- select --
Leased Car:	No agreement
Pension Fund:	-- select --
Promotion Possibilities:	-- select --
Working Hours:	9 hours
Your score:	294

Please write a complete, grammatically correct English sentence.

History of actions

#	Proposer	Action	Offer
3	Employer (You)	Message	I want to pay 12,000
4	Candidate	Message	I agree to { Salary : 12,000 NIS }. Now let's talk about the other issues. I offer you will be the project manager
5	Employer (You)	Message	I need QA
6	Candidate	Message	I cannot agree to { Job Description : QA }. I require Project Manager
7	Candidate	Message	my offer is project manager

Click here to quit the negotiation immediately:

Figure 1: The negotiation system's interface for NegoChat.

In order to improve the QOAgent, we extended it by using a generic opponent modeling mechanism, which allows the agent to model its counterpart's population and adapt its behavior to that population [32]. The extended agent, called KBAgent, is an automated negotiator that negotiates with each person only once, and uses past negotiation sessions of others as a knowledge base for generic opponent modeling. The database containing the a relatively small number of past negotiation sessions is used to extract the likelihood of acceptance of proposals and which proposals may be offered by the opposite side. The performance of KBAgent in terms of its counter-offer generation and generic opponent modeling was tested against people in the Tobacco and the Job interview domains.

The results of these tests indicate that the KBAgent negotiates proficiently with people and even achieves higher utility score values than the QOAgent. Moreover, the KBAgent achieves significantly better agreements, in terms of utility score, than the human counterparts in *both* roles. These results indicate that integrating general opponent modeling into qualtative decision-making is beneficial for automated negotiations.

2.3 NegoChat Agent

All the agents we discussed so far negotiated with the human counterpart either using a structured language or using a menu-driven interaction. They lack the natural language processing support required to enable real world types of interactions. To address this challenge we first developed an NLP module that translates the free text of the human player to the agent's formal language. We modified the KBagent by adding this module without changing the KBagent strategy and ran an experiment in which the modified KBagent played with people in a chat-like environment (see Figure 1 for the negotiation system's interface for chat-based negotiations). We found that simply modifying the KBagent to include an NLP module is insufficient to create a good agent for such settings and the revised agent achieved relatively low utility scores. The main observation was that people in chat-based negotiations make and accept partial agreements and follow issue-by-issue negotiations while the KBagent proposes full offers and has difficulties reaching partial agreements. To address this limitation, we developed NegoChat, which extended the KBagent focusing on strategies that allow for partial agreements and issue-by-issue interactions. NegoChat's algorithm is based on bounded rationality, specifically anchoring and Aspiration Adaptation Theory (AAT). The AAT was used for deciding on the order in which the issues will be discussed. The agent begins each negotiation interaction by proposing a full offer based on the KBagent's strategy, which serves as its anchor. Assuming this offer is not accepted, NegoChat then proceeds to negotiate via partial agreements, proposing the next issue for negotiation based on people's typical urgency (according to AAT).

We evaluated the NegoChat agent in extensive experiments negotiating with people in the job interview domain. We compared its performance to the performance of the KBAgent that also negotiated with (different) people using the same NLP module. The NegoChat agent achieved significantly better agreements (i.e., higher utility score) in less time. However, people playing against KBAgent, on average, did better. This implies that some of NegoChat's success is evidently at the cost of the person's score and consequently the social welfare score of this agent is not significantly better than that of KBAgent. As our goal is to maximize the agent's utility score this should not be seen as a fault. However, future generations of automated agents may decide to implement different strategies to maximize social welfare.

3. NEGOTIATIONS AND ACTIONS INTERLEAVING

In most situations, negotiation is not done in isolation but is associated with agreement implementation and other activities. We developed agents that can interact with people in such settings. These studies were carried out in a configurable system called Colored Trails (CT)[1]. It is a game played by two or more participants on a board of colored squares. CT is an abstract, conceptually simple but highly versatile game in which players negotiate and exchange resources to enable them to achieve their individual or group goals. It provides a realistic analogue to multi-agent task domains, while not requiring extensive domain modeling [15, 18]. A summary of the agents we developed are specified in Table 2.

3.1 Revelation games

We considered negotiation settings in which participants lack information about each other's preferences, often hindering their ability to reach beneficial agreements [34]. Specifically, we studied a particular class of such settings we call "revelation games", in which two players are given the choice to truthfully reveal private information before commencing two rounds of alternating negotiation. Revealing this information narrows the search space of possible agreements and may lead to agreement more quickly, but may also cause players to be exploited by others (see examples of such games in Figure 2). Revelation games combine two types of interactions that have been studied in the past in the economics literature: Signaling games [39], in which players choose whether to convey private information to each other, and bargaining [31], in which players engage in multiple negotiation rounds.

We were hopeful that, for revelation games, equilibrium-based agents will interact well with people since behavioral economics work has shown that people often follow equilibrium strategies [7] when deciding whether to reveal private information to others. The question is whether this observation will be stronger than our previous observations reported above that people's behavior in bargaining settings does not adhere to equilibrium strategies. We formalized the setting as a Bayesian game and computed two types of perfect Bayesian equilibrium: a separating equilibrium where both players reveal their type, and a pooling equilibrium where none of the players reveal their types.

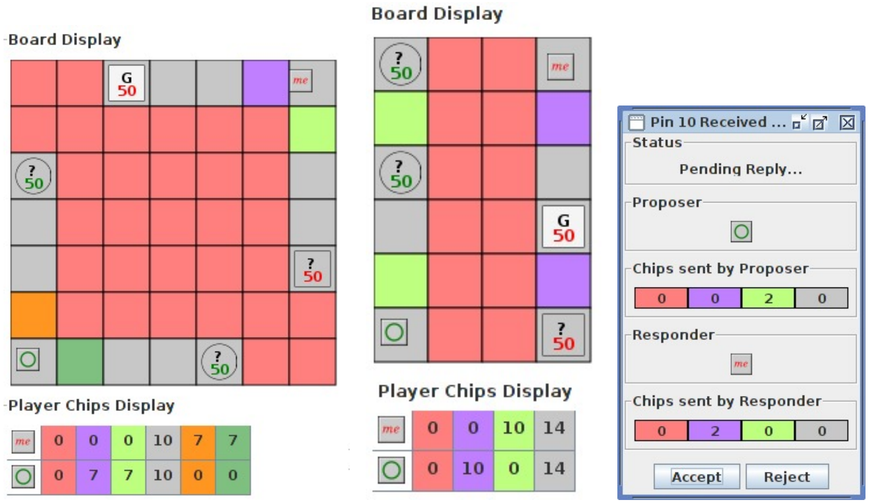
We compared the equilibrium agents with people playing with other people and with the Sigmoid Acceptance Learning Agent (SIGAL) that we developed [34]. The SIGAL agent used classical machine learning techniques to predict how people make and respond to offers during negotiation, how they reveal information and their response to potential revelation actions by the agent. This model is integrated into the agent's decision tree. We conducted an extensive empirical study spanning hundreds of human subjects.

Results show that the SIGAL agent was able to outperform people and the equilibrium agents. Furthermore, people outperformed the equilibrium agents. It turned out that the negotiation part of the game was more important (with respect to the utility score) than the revelation part. The equilibrium agent made very selfish offers in the last round of the negotiations. Most of these offers were rejected. In the first round, it made offers that were highly beneficial to people and most of these offers were accepted, but the small benefit it incurred in these proposals did not aid its performance. The SIGAL agent, on the other hand, (i) learned to make offers that were beneficial to people while not compromising its own benefit; and (ii) incrementally revealed information to people in a way that increased its expected performance. We were able to adjust SIGAL to new, similar settings that varied rules and situational parameters of the game without the need to accumulate new data. However, moving to a completely new setting requires a lot of work collecting data and adjusting the machine learning module to the new setting.

3.2 Non-binding agreements

We also studied CT settings of two players in which both participants needed to complete their individual tasks by reaching agreements and exchanging resources, the number of negotiation rounds were not fixed in advance, and the negotiation protocol was an alternating offers protocol that allowed parties to choose the extent to which they kept each of their agreements during the negotiation [19]. That is, there are three phases in each round of the game: negotiation, transfer and movement. The negotiation phase consisted of two rounds of alternating offers in which the players could reach an agreement on resource exchange. After each phase of negotiations, the game moved to the "transfer phase" in

[1] See http://www.eecs.harvard.edu/ai/ct.



me	0	0	0	10	7	7
○	0	7	7	10	0	0

me	0	0	10	14
○	0	10	0	14

Status: Pending Reply...

Proposer: ○

Chips sent by Proposer: 0 0 2 0

Responder: me

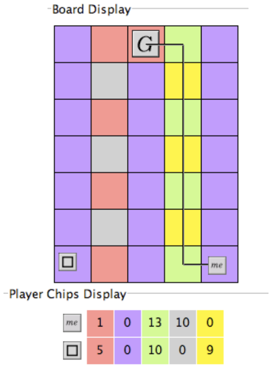
Chips sent by Responder: 0 2 0 0

(a) Symmetric Board Game (b) Asymmetric Board Game (c) A possible proposal

Figure 2: Two CT revelation games

Table 2: CT games

Settings	Agent Name	Agent Properties	Significance vs people
Bilateral, Uncertainty, Revelation Game two-phases: revelation, bargaining	PBE agent	Bayesian perfect equilibrium	No roles
	SIGAL	decision theory, machine learning	Both roles
Bilateral, full information, agreements not enforceable, multiple rounds, three phases: bargaining, resource exchange, movement	PAL	decision theory: Influence Diagram machine learning	Both roles
Contract game, three players two-phases: bargaining, movement	SPE agent	subgame-perfect equilibrium	CS role
	SP-RAP	subgame-perfect equilibrium bounded rational model of opponent risk averse	SP role



me	1	0	13	10	0
□	5	0	10	0	9

Figure 3: An example of a CT Board for multiple negotiation games with unenforceable agreements.

which both players could transfer resources to each other. The transfer action was done simultaneously, such that neither player could see what the other player transferred until the end of the phase. A player could choose to transfer more resources than it agreed to, or any subset of the resources it agreed to, including not transferring any resources at all. In the "movement phase" both players could move their icons on the board one step towards the goal square, provided they had the necessary resources. Then, the game moved to the next round, beginning again with negotiation phase. The game ends when one of the players reaches his goal or does not move for two rounds (see an example of one such game in Figure 3).

The most important decision of a player in such settings is whether or not to keep the agreements. Another important decision is whether to accept an offer given by the other player. In subgame perfect equilibrium, the players should not keep the agreements. Different equilibria may specify various strategies for the acceptance decision. We ran preliminary experiments and observed that such strategies are not beneficial when the equilibrium agent interacts with people. Most of the time the agent was not able to reach its goal, yielding a low utility score.

Galit et al. [19] present the Personality Adaptive Learning (PAL) agent for negotiating with people from different cultures for the CT game where agreements are not enforceable. The methodology was similar to that of SIGAL (Section 3.1), combining a decision-theoretic model using a decision tree with classical machine learning techniques to predict how people respond to offers, and the extent to which they fulfill agreements.

PAL was evaluated empirically in the Colored Trails (CT)

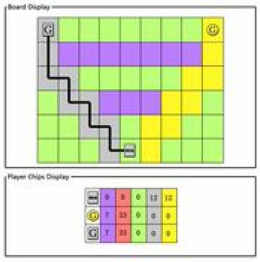


Figure 4: An example of a CT Board for the Contract game.

environment by playing with people in three countries: Lebanon, the U.S., and Israel, in which people are known to vary widely in their negotiation behavior. The agent was able to outperform people in all three countries.

3.3 Contract Game

We studied commitment strategies in a three-player CT game. The game is called Contract Game and is analogous to a market setting in which participants need to reach agreements over contracts and commit to or renege from contracts over time in order to succeed [20]. The game comprises three players, two service providers and one customer. The service providers compete to make repeated contract offers to the customer consisting of resource exchanges in the game (see an example of one such game in Figure 4). We formally analyzed the game to compute subgame perfect equilibrium strategies for the customer and service provider in the game that are based on making contracts containing commitment offers. To evaluate agents that use the equilibrium strategies, we conducted extensive empirical studies in three different countries, the U.S., Israel and China. We ran several configurations in which two human participants played a single agent participant in various role configurations in the game. Our results showed that the computer agent using subgame Nash equilibrium strategies for the customer role was able to outperform people playing the same role in all three countries and obtained statistically significant, higher utility scores than the humans. This was very surprising since it was the first EQ agent after trying many equilibrium agents that was able to achieve such results.

In particular, the customer agent made significantly more commitment type proposals than people did, and requested significantly more resources from service providers than did people. It was quite surprising that people playing with it accepted these offers; in other settings (such as the revelation games) such unfair offers were rejected by people. We hypothesize that the competition between the two service providers made such offers more acceptable. In addition, while in the revelation games the EQ agent had only one opportunity to make an offer, in the contract game it could make offers several times (off the equilibrium path) which we believe also increased the acceptance rate. Also, the customer agent reached one of the goals in all its games and was able to reach the goal significantly more often than people. This is again quite surprising since at the beginning of the game the customer has enough resources to reach both goals. Since reaching one of the goals is very beneficial to the customer it is difficult to understand why human players hadn't always reached the goal.

While the customer EQ agent outperformed people, people outperformed the EQ agent when it played the role of one of the service providers. We believe that this is mainly due to people playing the customer role not reaching the goal even when they have all the needed resources to do so. To face this problem we then developed an agent termed SP-RAP which extended the EQ agent in the following two ways to handle the uncertainty that characterizes human play in negotiation: First, it employed a risk averse strategy using a convex utility function. Second, it reasoned about a possibly bounded rational customer (CS) player by assigning a positive probability $p > 0$ for the customer player not reaching the goal. We assigned a separate value for p for each country by dividing the number of times the CS player reached the goal by the total number of games played. Consequently, SP-RAP outperformed people playing the SP role in all three countries.

4. ARGUMENTATION AGENT

An automated agent can help a human when engaging in an argumentative dialog by utilizing its knowledge and computational advantage to provide arguments to him. Argumentation was studied extensively using the well-established Argumentation Theory (see [41] for a summary). Therefore, in the first step in the development of an automated agent that advised people in such settings, we considered the abilities of Argumentation Theory to predict people's arguments. In [38] we presented extensive studies in three experimental settings, varying in complexity, which show the lack of predictive power of the existing Argumentation Theory. Second, we used Machine Learning (ML) techniques to provide a probability distribution over all known arguments given a partial deliberation. That is, our ML techniques provided the probability of each argument to be used next in a given dialog. Our model achieves 76% accuracy when predicting people's top three argument choices given a partial deliberation. Last, using the prediction model and the newly introduced heuristics of relevance, we designed and evaluated the Predictive and Relevance based Heuristic agent (PRH). Through an extensive human study, we showed that the PRH agent outperforms other agents that propose arguments based on Argumentation Theory, predicted arguments without heuristics or only the heuristics on both axes we examined: people's satisfaction from agents and people's use of the suggested arguments.

5. SECURITY GAMES

The last several years have witnessed the successful application of Bayesian Stackelberg games in allocating limited resources to protect critical infrastructures. These interesting efforts have been led by Prof. Milind Tambe from USC. The first application is the ARMOR system (Assistant for Randomized Monitoring over Routes) that has been deployed at the Los Angeles International Airport (LAX) since 2007 to randomize checkpoints on the roadways entering the airport and canine patrol routes within the airport terminals [33, 35]. Other applications include IRIS, a game-theoretic scheduler for randomized deployment of the US Federal Air Marshal Service (FAMS) requiring significant scale-up in underlying algorithms, which has been in use since 2009 [40]; and PROTECT, which requires further scale up, is deployed for generating randomized patrol schedules for the US Coast Guard in Boston, New York, Los Angeles and other ports around the US [2, 3]. Furthermore, TRUSTS has been evaluated for deterring fare evasion, suppressing urban crime and counter-terrorism within the Los Angeles Metro System [44, 21, 9] and GUARDS was earlier tested by the US Transportation Security Administration (TSA) for security inside the airport [37].

The evaluation of these systems could be quite limited. The only system that was truly evaluated in the field is TRUSTS. We were able to conduct controlled experiments of our game theoretic resource allocation algorithms. Before this project, the actual evaluation of the deployed security games applications in the field was a major open challenge. The reasons were twofold. First, previous applications focused on counter-terrorism, therefore controlled experiments against real adversaries in the field were not feasible. Second, the number of practical constraints related to real-world deployments limited the ability of researchers to conduct head-to-head comparisons

In TRUSTS we were able to address this challenge and run the largest scale evaluation of security games in the field in terms of duration and number of security officials deployed. We evaluated each component of the system (Fare Evasion, Counter Terrorism and Crime algorithms) by designing and running field experiments. In the context of fare evasion, we ran an extensive experiment, where we compared schedules generated using game theory against competing schedules comprised of a random scheduler, augmented with officers providing real-time knowledge of the current situation. Our results showed that our schedules led to statistically significant improvements over the competing schedules, despite the fact that the latter were improved with real-time knowledge.

In addition, extensive human experiments in the lab were conducted [36, 30]. These experiments showed that incorporating bounded rational models of the adversaries to the Stackelberg games improves the performance of the defenders. These results were observed both when the role of adversaries was played by novices and when it was played by security experts.

6. PERSUASION GAMES

A persuasion game involves two players: a sender who attempts to persuade another agent (the receiver) to take a certain action [17]. Persuasion games are similar to both negotiation games and security games [43]. They are similar to negotiation and argumentation games since one player tries to convince another player to do something, as in negotiations. They are also similar to security games in the asymmetry between the players: the sender and the defender are trying to influence the activities of the receiver and the attacker, respectively. So, it is interesting to check if equilibrium strategies will be beneficial in persuasion games. Furthermore, the incorporation of a bounded rational model of the receiver will be beneficial to the sender as the incorporation of bounded rational models of the attacker was beneficial to the defender in security games.

We focused on information disclosure games with two-sided uncertainty [5, 6]. This is a special type of persuasion game in which an agent tries to lead a person to take an action that is beneficial to the agent by providing him with truthful, but possibly partial, information relevant to the action selection. We first computed the subgame perfect Bayesian Nash equilibrium of the game assuming the human receiver is fully rational. We developed a sender agent that follows the equilibrium strategy (GTBA agent).

We also developed a machine learning-based model that effectively predicts people's behavior in these games and we called it Linear weighted-Utility Quantal response (LUQ). The model we provide assumes that people use a subjective utility function which is a linear combination for all given attributes. The model also assumes that while people use this function as a guideline, they do not always choose the action with the greatest utility value, however, the higher an action's utility value is, the more likely they are to choose that action. We integrated this model into our persuasion model and built the LUQA agent.

We ran an extensive empirical study with people in two different games. In a multi-attribute road selection game with two-sided uncertainty, the LUQA agent obtained significantly higher utility points than the GTBA agent. However, in the second game, the Sandwich game, there was no significant advantage to the machine learning-based model, and using the game theory-based agent, GTBA, which assumes that people maximize their expected monetary values is beneficial. We hypothesize that these different results are due to the nature of the domains. The monetary result plays an important role in the sandwich game. This is because the game is played in an environment where a person's goal is to make a profit. However, in the road selection game the utility scores are associated with time. Thus, it seems that maximizing expected monetary utility is easier for people than maximizing utility scores that are associated with time.

7. DISCUSSIONS

The state-of-the-art agent, NegoChat, for multi-issue negotiations integrates methods from several disciplines: qualitative decision-making, machine learning and heuristics based on psychological theories. None of the equilibrium agents that were developed were successful when interacting with people. The reliance on heuristic and machine learning makes the transfer of NegoChat from one scenario to the other and from one culture to the other problematic. This was evident recently when we tried to run experiments with NegoChat, which was developed based on data collected in Israel and Egypt. We had to spend a lot of time and effort until this transfer was possible.

Similarly, in most of the cases, the equilibrium agent was not successful in the CT game settings. The only exception is the contract game. We believe that the success of the equilibrium agent in the contract game has to do with the specifics of the game: the competition between the two SPs. In the contract game, it was extremely difficult to predict people's behavior, thus the success of the equilibrium agent is even more significant.

The same observations were seen in argumentation – the argumentation theory-based agent was not very successful. Therefore, in all these cases the development of new negotiation agents to new settings requires the collection of data and the adjustment of the agent to the new settings. Therefore, we strongly believe that the development of theoretical models for the design and implementation of agents that negotiate in multi-issue negotiation settings can be very useful. However, this is still an open question.

On the other hand, it seems that in security games the deployment of Stackelberg equilibrium is beneficial (possibly with the incorporation of a bounded rational model of the attacker) and similarly in persuasion games where using subgame perfect Baysian equilibrium is beneficial (possibly with the incorporation of a bounded rational model of the receiver).

We hypothesize that this is the case since in security games and persuasion games the interactions between the agent and the human is quite limited. The attacker or the receiver needs to choose one action compared to many decision-making activities that are required from a human negotiator. Nevertheless, even in security games and to some extent in persuasion games it was shown that taking the limitations of the other player into consideration is beneficial.

8. ACKNOWLEDGMENT

This work was supported in part by ERC grant #267523.

9. REFERENCES

[1] O. Amir, B. Grosz, E. Law, and R. Stern. Collaborative health care plan support. In *Proceedings of the Eleenth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2013)*, pages 793–796, St Paul, MN, 2013.

[2] B. An, M. Jain, M. Tambe, and C. Kiekintveld. Mixed-initiative optimization in security games: A preliminary report. In *AAAI Spring Symposium: Help me help you: Bridging the gaps in human-agent collaboration*, 2011.

[3] B. An, F. Ordóñez, M. Tambe, E. Shieh, R. Yang, C. Baldwin, J. DiRenzo III, K. Moretti, B. Maule, and G. Meyer. A deployed quantal response-based patrol planning system for the us coast guard. *Interfaces*, 43(5):400–420, 2013.

[4] D. Ariely. *Predictably Irrational*. Harper Collins, 2008.

[5] A. Azaria, Z. Rabinovich, C. V. Goldman, and S. Kraus. Strategic information disclosure to people with multiple alternatives. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):64, 2014.

[6] A. Azaria, Z. Rabinovich, S. Kraus, and C. V. Goldman. Strategic information disclosure to people with multiple alternatives. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*, 2011.

[7] J. Banks, C. F. Camerer, and D. Porter. Experimental tests of Nash refinements in signaling games. *Games and Economic Behavior*, 6:1–31, 1994.

[8] M. H. Bazerman and M. A. Neale. Negotiator rationality and negotiator cognition: The interactive roles of prescriptive and descriptive research. In H. P. Young, editor, *Negotiation Analysis*, pages 109–130. The University of Michigan Press, 1992.

[9] F. M. Delle Fave, A. X. Jiang, Z. Yin, C. Zhang, M. Tambe, S. Kraus, and J. P. Sullivan. Game-theoretic patrolling with dynamic execution uncertainty and a case study on a real transit system. *Journal of Artificial Intelligence Research*, pages 321–367, 2014.

[10] E. Durenard. *Professional Automated Trading: Theory and Practice*. John Wiley & Sons, 2013.

[11] A. Elmalech, D. Sarne, and B. J. Grosz. Problem restructuring for better decision making in recurring decision situations. *Autonomous Agents and Multi-Agent Systems*, 29(1):1–39, 2015.

[12] A. Elmalech, D. Sarne, A. Rosenfeld, and E. S. Erez. When suboptimal rules. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1313–1319, 2015.

[13] I. Erev and A. E. Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, 88(4):848–881, 1998.

[14] S. Fatima, S. Kraus, and M. Wooldridge. *Principles of Automated Negotiation*. Cambridge University Press, 2014.

[15] Y. Gal, B. Grosz, S. Kraus, A. Pfeffer, and S. Shieber. Agent decision-making in open mixed networks. *Artificial Intelligence*, 174(18):1460–1480, 2010.

[16] A. D. Galinsky and T. Mussweiler. First offers as anchors: the role of perspective-taking and negotiator focus. *Journal of personality and social psychology*, 81(4):657, 2001.

[17] J. Glazer and A. Rubinstein. A study in the pragmatics of persuasion: a game theoretical approach. *Theoretical Economics*, 1(4):395–410, 2006.

[18] B. J. Grosz, S. Kraus, S. Talman, B. Stossel, and M. Havlin. The influence of social dependencies on decision-making: Initial investigations with a new game. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 782–789. IEEE Computer Society, 2004.

[19] G. Haim, Y. Gal, M. Gelfand, and S. Kraus. A cultural sensitive agent for human-computer negotiation. In *Proceedings of the Eleventh International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2012)*, pages 451–458, Valencia, Spain, 2012.

[20] G. Haim, Y. K. Gal, S. Kraus, and B. An. Human-computer negotiation in three-player market settings. In *Proc. of ECAI14*, pages 417–422, 2014.

[21] A. X. Jiang, Z. Yin, C. Zhang, M. Tambe, and S. Kraus. Game-theoretic randomization for security patrolling with dynamic execution uncertainty. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 207–214. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[22] K. Kauppi, A. Brandon-Jones, S. Ronchi, and E. van Raaij. Tools without skills: Exploring the moderating effect of absorptive capacity on the relationship between e-purchasing tools and category performance. *International Journal of Operations & Production Management*, 33(7):828–857, 2013.

[23] G. Kersten and H. Lai. Negotiation support and e-negotiation systems: an overview. *Group Decision and Negotiation*, 16(6):553–586, 2007.

[24] S. Kraus, P. Hoz-Weiss, J. Wilkenfeld, D. R. Andersen, and A. Pate. Resolving crises through automated bilateral negotiations. *Artificial Intelligence*, 172(1):1–18, 2008.

[25] D. A. Lax and J. K. Sebenius. Thinking coalitionally:

party arithmetic, process opportunism, and strategic sequencing. In H. P. Young, editor, *Negotiation Analysis*, pages 153–193. The University of Michigan Press, 1992.

[26] R. Lin, S. Kraus, J. Wilkenfeld, and J. Barry. Negotiating with bounded rational agents in environments with incomplete information using an automated agent. *Artificial Intelligence*, 172(6):823–851, 2008.

[27] R. Lin, Y. Oshrat, and S. Kraus. Investigating the benefits of automated negotiations in enhancing people's negotiation skills. In *Proceedings of the Eighth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2009)*, pages 345–352, Budapest, Hungary, 2009.

[28] M. Maschler, E. Solan, and S. Zamir. *Game Theory*. Cambridge University Press: Cambridge, England, 2013.

[29] R. D. McKelvey and T. R. Palfrey. An experimental study of the centipede game. *Econometrica*, 60(4):803–836, 1992.

[30] T. H. Nguyen, R. Yang, A. Azaria, S. Kraus, and M. Tambe. Analyzing the effectiveness of adversary modeling in security games. In *Proc of AAAI 2013*, 2013.

[31] M. J. Osborne and A. Rubinstein. *A course in game theory*. MIT press, 1994.

[32] Y. Oshrat, R. Lin, and S. Kraus. Facing the challenge of human-agent negotiations via effective general opponent modeling. In *Proceedings of the Eighth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2009)*, pages 377–384, Budapest, Hungary, 2009.

[33] P. Paruchuri, J. P. Pearce, J. Marecki, M. Tambe, F. Ordonez, and S. Kraus. Playing games for security: an efficient exact algorithm for solving bayesian stackelberg games. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*, pages 895–902. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[34] N. Peled, K. Gal, and S. Kraus. A study of computational and human strategies in revelation games. *Autonomous Agents and Multi-Agent Systems*, 29(1):73–97, 2015.

[35] J. Pita, M. Jain, J. Marecki, F. Ordóñez, C. Portway, M. Tambe, C. Western, P. Paruchuri, and S. Kraus. Deployed armor protection: the application of a game theoretic model for security at the los angeles international airport. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems: industrial track*, pages 125–132. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[36] J. Pita, M. Jain, M. Tambe, F. Ordóñez, and S. Kraus. Robust solutions to stackelberg games: Addressing bounded rationality and limited observations in human cognition. *Artificial Intelligence*, 174(15):1142–1171, 2010.

[37] J. Pita, M. Tambe, C. Kiekintveld, S. Cullen, and E. Steigerwald. Guards - innovative application of game theory for national airport security. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, page 2710, 2011.

[38] A. Rosenfeld and S. Kraus. Providing arguments in discussions based on the prediction of human argumentative behavior. In *Proc. of AAAI*, 2015.

[39] A. M. Spence. *Market signaling: Informational transfer in hiring and related screening processes*, volume 143. Harvard Univ Pr, 1974.

[40] J. Tsai, C. Kiekintveld, F. Ordonez, M. Tambe, and S. Rathi. Iris-a tool for strategic security allocation in transportation networks. In *Proc. of AAMAS09*, 2009.

[41] D. Walton. Argumentation theory: A very short introduction. In *Argumentation in artificial intelligence*, pages 1–22. Springer, 2009.

[42] J. Wilkenfeld, S. Kraus, K. M. Holley, and M. A. Harris. Genie: A decision support system for crisis negotiations. *Decision Support Systems*, 14(4):369–391, 1995.

[43] H. Xu, Z. Rabinovich, S. Dughmi, and M. Tambe. Exploring information asymmetry in two-stage security games. In *Proc of AAAI*, 2015.

[44] Z. Yin, A. X. Jiang, M. Tambe, C. Kiekintveld, K. Leyton-Brown, T. Sandholm, and J. P. Sullivan. Trusts: Scheduling randomized patrols for fare inspection in transit systems using game theory. *AI Magazine*, 33(4):59, 2012.

Rationality and Beliefs in Dynamic Games: Revealed-Preference Foundations

[Extended Abstract]

Marciano Siniscalchi
Economics Department, Northwestern University
2001 Sheridan Rd
Evanston, IL 60611, USA
marciano@northwestern.edu

General Terms

Game Theory

Keywords

Conditional Probability Systems, Decision Theory, Game Theory, Sequential Rationality

1. EXTENDED ABSTRACT

Sequential rationality is the prevalent notion of best response for dynamic games; it is an essential part of the definition of sequential equilibrium [9], perfect Bayesian equilibrium [6], and extensive-form rationalizability [11]. Abstracting from notational and other minor differences, a strategy s_i of player i is sequentially rational if, beginning at any information set where i moves, s_i specifies a sequence of actions that is optimal given the beliefs that i holds at that information set about the play of her opponents.

While this notion is central to the theory of dynamic games, it raises both practical and methodological concerns. From a practical standpoint, it is not obvious how to reliably ascertain which strategy a player follows in a given dynamic game, and a fortiori whether it is or isn't sequentially rational. Consider a three-stage Centipede game [13], played by Ann (who moves at the first and third nodes) and Bob. Suppose that, as predicted by backward induction, Ann ends the game at the first node by moving Down. Then, we are unable to observe Bob's intended choice at the second node. Furthermore, to determine whether such a choice would have been optimal, had Ann chosen Across instead, we need to consider Bob's beliefs conditional upon an event that Bob does not expect to occur—a zero-probability event. While formally we can represent such beliefs, it is not clear how an experimenter might elicit them in practice.

Reinhard Selten's *strategy method* [16] is a widely used experimental procedure that is intended to elicit players' intended strategy choices in dynamic games. There is evidence that the strategy method is an effective elicitation procedure: see, e.g., [3]. But this finding actually raises further questions. The strategy method essentially asks players to simultaneously *commit* to a strategy, which is then implemented by the experimenter without possibility of subsequent intervention by the subjects. This reduces the original dynamic game to one in which players face non-trivial moves only in the initial stage; furthermore, such moves are simultaneous. Standard solution concepts such as sequential or perfect Bayesian equilibrium predict that players will maximize *ex-ante* expected payoffs when the strategy method is employed; therefore, such solution concepts do allow subjects to commit to strategies that are not sequentially rational in the original game. Refinements that incorporate the notion of *invariance* [8] do imply that subjects will only commit to strategies that are sequentially rational in the original game. However, there is ample experimental evidence that contradicts the invariance hypothesis [5, 15, 4, 7]. Thus, the received theory cannot at the same time explain the effectiveness of the strategy method, and account for violations of the invariance hypothesis.

These practical issues hint at a deeper methodological concern. Economics has long embraced the revealed-preference approach: assumptions about agents' tastes and beliefs should be testable, or elicitable, on the basis of observable choices in suitably designed problems. To date, rationality and beliefs in dynamic games have not been subject to analysis from the revealed-preference perspective. Formal definitions of sequential rationality build upon expected-utility maximization. However, the revealed-preference foundations for expected utility [14, 1] concern atemporal, or one-shot choices. Furthermore, extensions of expected-utility theory to dynamic choice problems are wholly silent about behavior conditional upon ex-ante zero probability events. Of course, the analysis of intended choices following unexpected moves is at the heart of dynamic game theory. Thus, the received decision theory is insufficient to provide foundations to the analysis of dynamic games.

The objective of this project is to provide such a foundation. This entails two contributions. The first is to define a novel choice criterion for dynamic decision problems and games, *sequential preference*, so as to satify two criteria. First, the proposed criterion implies sequential rationality. Second, it allows preferences over strategies to be elicited from ex-ante choices, using a version of the strategy method. In particular, sequential preferences provide a theoretical rationale for the use of this common experimental procedure, as well as a method to elicit conditional beliefs following zero-probability events.

Building on the finding that sequential preferences are indeed elicitable, the second contribution of this project is to provide a behavioral, or axiomatic, characterization of the proposed choice criterion. This is based on a suitable adaptation of the Anscombe-Aumann [1] axioms.

Finally, in the analysis of sequential preferences, there are natural connections to conditional probability systems [12, 10] and lexicographic probability systems [2]. These are explored, and their implications for game-theoretic analysis are discussed.

The project is carried out in two papers. The first, [18], introduces sequential preferences, and analyzes elicitation and the strategy method. The second, [17], provides behavioral foundations.

2. REFERENCES

[1] F. J. Anscombe and R. J. Aumann. A definition of subjective probability. *Annals of Mathematical Statistics*, 34:199–205, 1963.

[2] L. Blume, A. Branderburger, and E. Dekel. Lexicographic probabilities and choice under uncertainty. *Econometrica*, 59:61–79, 1991.

[3] J. Brandts and G. Charness. The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, 14(3):375–398, 2011.

[4] D. J. Cooper and J. B. Van Huyck. Evidence on the equivalence of the strategic and extensive form representation of games. *Journal of Economic Theory*, 110(2):290–308, 2003.

[5] R. Cooper, D. V. DeJong, R. Forsythe, and T. W. Ross. Forward induction in the battle-of-the-sexes games. *American Economic Review*, 83(5):1303–1316, 1993.

[6] D. Fudenberg and J. Tirole. Perfect Bayesian equilibrium and sequential equilibrium* 1. *Journal of Economic Theory*, 53(2):236–260, 1991.

[7] S. Huck and W. Müller. Burning money and (pseudo) first-mover advantages: an experimental study on forward induction. *Games and Economic Behavior*, 51(1):109–127, 2005.

[8] E. Kohlberg and J. Mertens. On the strategic stability of equilibria. *Econometrica: Journal of the Econometric Society*, 54(5):1003–1037, 1986.

[9] D. Kreps and R. Wilson. Sequential equilibria. *Econometrica: Journal of the Econometric Society*, 50(4):863–894, 1982.

[10] R. B. Myerson. Axiomatic foundations of bayesian decision theory. Discussion Paper 671, The Center for Mathematical Studies in Economics and Management Science, Northwestern University, January 1986.

[11] D. G. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52:1029–1050, 1984.

[12] A. Rényi. On a new axiomatic theory of probability. *Acta Mathematica Hungarica*, 6(3):285–335, 1955.

[13] R. Rosenthal. Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox. *Journal of Economic Theory*, 25(1):92–100, 1981.

[14] L. Savage. *The foundations of statistics.* Dover Pubns, 1972.

[15] A. Schotter, K. Weigelt, and C. Wilson. A laboratory investigation of multiperson rationality and presentation effects. *Games and Economic behavior*, 6(3):445–468, 1994.

[16] R. Selten. Ein oligopolexperiment mit preisvariation und investition. *Beiträge zur experimentellen Wirtschaftsforschung, ed. by H. Sauermann, JCB Mohr (Paul Siebeck), Tübingen*, pages 103–135, 1967.

[17] M. Siniscalchi. Foundations for sequential preferences. mimeo, Northwestern University, 2015.

[18] M. Siniscalchi. Sequential preferences and sequential rationality. mimeo, Northwestern University, 2015.

Recent Methodological Advances in Causal Discovery and Inference

[Invited Paper]

Peter Spirtes
Department of Philosophy
Carnegie Mellon University
ps7z@andrew.cmu.edu

Kun Zhang
MPI for Intelligent Systems
72076 Tübingen, Germany &
Info. Sci. Inst., USC
4676 Admiralty Way, CA 90292
kzhang@tuebingen.mpg.de

ABSTRACT

This paper aims to give a broad coverage of central concepts and principles involved in automated causal inference and emerging approaches to causal discovery from i.i.d data and from time series. After reviewing concepts including manipulations, causal models, sample predictive modeling, causal predictive modeling, and structural equation models, we present the constraint-based approach to causal discovery, which relies on the conditional independence relationships in the data, and discuss the assumptions underlying its validity. We then focus on causal discovery based on structural equations models, in which a key issue is the identifiability of the causal structure implied by appropriately defined structural equation models: in the two-variable case, under what conditions (and why) is the causal direction between the two variables identifiable? We show that the independence between the error term and causes, together with appropriate structural constraints on the structural equation, makes it possible. Next, we report some recent advances in causal discovery from time series. Assuming that the causal relations are linear with non-Gaussian noise, we study two problems which are traditionally difficult to solve, namely, causal discovery from subsampled data and that in the presence of confounding time series. Finally, we list a number of open questions in the field of causal discovery and inference.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Miscellaneous*

General Terms

Algorithms, Theory

Keywords

Causal inference, causal discovery, structural equation model, conditional independence, statistical independence, identifiability

1. INTRODUCTION

The goal of many sciences is to understand the mechanisms by which variables came to take on the values they have (i.e., to find a generative model), and to predict what the values of those variables would be if the naturally occurring mechanisms in a population[1] were subject to outside *manipulations*. For example, a *randomized experiment* is one kind of manipulation, which substitutes the outcome of a randomizing device to set the value of a variable, such as whether or not a particular diet is used, instead of the naturally occurring mechanism that determines diet. In non-experimental settings, biologists gather data about the gene activation levels in normally operating systems, and seek to understand which genes affect the activation levels of which other genes, and seek to predict what the effects of intervening to turn some genes on or off would be; epidemiologists gather data about dietary habits and life expectancy in the general population and seek to find what dietary factors affect life expectancy and to predict the effects of advising people to change their diets. Finding answers to questions about the mechanisms by which variables come to take on values, or predicting the value of a variable after some other variable has been manipulated, is characteristic of causal inference. If only observational (non-experimental) data is available, predicting the effects of manipulations typically involves drawing samples from one density (of the unmanipulated population) and making inferences about the values of a variable in a population that has a different density (of the manipulation population).

Many of the basic problems and basic assumptions remain the same across domains. In addition, although there are some superficial similarities between traditional supervised machine learning problems and causal inference (e.g., both employ model search and feature selection, the kinds of models employed overlap, some model scores can be used for both purposes), these similarities can mask some very important differences between the two kinds of problems.

1.1 History

Traditionally, there have been a number of different approaches to causal discovery. The gold standard of causal discovery has typically been to perform planned or randomized experiments [10]. There are obvious practical and ethical considerations that limit the application of randomized

[1] Here, the "population" is simply a collection of instantiations of a set of random variables. For example, it could consist of a set of satellite readings and rainfall rates in different locations at a given time, or the readings of a single satellite and rainfall rate over time, or a combination of these.

experiments in many instances, particularly on human beings. Moreover, recent data collection techniques and causal inference problems raise several practical difficulties regarding the number of experiments that need to be performed in order to answer all of the outstanding questions [8, 9].

2. MANIPULATING AND CONDITIONING

Conditioning maps a given joint density, and a given subpopulation (typically specified by a set of values for random variables) into a new density. The conditional density is a function of the joint density over the random variables, and a set of values for a set of random variables.[2] The estimation of a conditional probability is often non-trivial because the number of measurements in which the variables conditioned on that take on a particular value might be small. A large part of statistics and machine learning is devoted to estimating conditional probabilities from realistic sample sizes under a variety of assumptions.

More generally, suppose the goal is to find a "good" predictor of the value of some target variable Y from the values of the observed covariates $\mathbf{O}$, for a unit. We will refer to this as Problem 1, described more formally below. Ultimately, the prediction of the value of Y is performed by some prediction function $\hat{Y}_n(\mathbf{O})$. One traditional measure of how good the predictor $\hat{Y}_n(\mathbf{O})$ is in predicting Y is the mean squared prediction error (MSPE), which is equal to $E[(Y-\hat{Y}_n(\mathbf{O}))^2]$, where the expected value is taken with respect to the density $p(\mathbf{O}, Y)$ [1].[3]

Problem 1: Sample predictive modeling

Input: i.i.d. samples from a population with density $p(\mathbf{O}, Y)$, background assumptions, and a target variable Y whose value is to be predicted.

Output: $\hat{Y}_n(\mathbf{O})$, a predictor of Y from $\mathbf{O}$ that has a small MSPE.

In addition to predicting future values of random variables from the present and past values, conditional probabilities are also useful for predicting hidden values at the current time.

2.1 Manipulated Probabilities

A *manipulated* density results from taking action on a given population – it may or may not be equal to any observational conditional density, depending upon what the causal relations between variables are. Manipulated probability densities are the appropriate probability densities to use when making predictions about the effects of taking actions ("manipulating" or "doing") on a give population (e.g., assigning satellite readings), rather than observing ("seeing") the values of given variables. A manipulation M specifies a new conditional probability density for some set of variables. If $\mathbf{X}$ and $\mathbf{O}$ are sets of variables with density $p(\mathbf{X}|\mathbf{O})$, a manipulation M changes the density to some new density $p'(\mathbf{X}|\mathbf{O})$. Manipulated probabilities are the probabilities that are implicitly used in decision theory, where the different actions under consideration are manipulations.[4] We designate the density of a set of variables $\mathbf{V}$ after a manipulation M as $p(\mathbf{V}||M)$. Each manipulation is assumed to be an ideal manipulation in the following senses:

1. Each manipulation succeeds, i.e., if the manipulation is designated as setting the density to $p'(\mathbf{X}|\mathbf{O})$, then the post-manipulation density is $p'(\mathbf{X}|\mathbf{O})$.

2. There is no fat hand, i.e., each manipulation directly affects only the variables manipulated.

A probability model specifies a density over a set of random variables $\mathbf{O}$. A causal model specifies a set of densities over a set of random variables $\mathbf{O}$, one for each possible manipulation M of the random variables in $\mathbf{O}$, including the null manipulation. Hence a probability model is a member of a causal model.

Given a set of variables $\mathbf{V}$, the direct causal relations among the variables can be represented by a directed graph, where the variables in $\mathbf{V}$ are the vertices, and there is an edge from A to B if A is a direct cause of B relative to $\mathbf{V}$.

We will refer to the problem of estimating manipulated densities given a sample from a marginal unmanipulated density, a (possibly empty) set of samples from manipulated densities, and background assumptions, as Problem 2; it is stated more formally below. In contrast to conditional probabilities, which can be estimated from samples from a population, typically the gold standard for estimating manipulated densities is an experiment, often a randomized trial. However, in many cases experiments are too expensive, too difficult, or not ethical to carry out. This raises the question of what can be determined about manipulated probability densities from samples from a population, possibly in combination with a limited number of randomized trials. The problem is even more difficult because the inference is made from a set of measured random variables $\mathbf{O}$ from samples that might not contain variables that are causes of multiple variables in $\mathbf{O}$.

Problem 2 is usually broken into two parts: finding a set of causal models from sample data, some manipulations (experiments) and background assumptions, and predicting the effects of a manipulation given a causal model. Here, a "causal model" (Section 3) specifies for each possible manipulation that can be performed on the population (including the manipulation that does nothing to a population) a post-manipulation density over a given set of variables.

Problem 2: Statistical causal predictive modeling

Input: i.i.d. samples from a population with density $p(\mathbf{O}, Y)$, a (possibly empty) set of i.i.d. samples from manipulated densities $p(\mathbf{O}, Y||M_1), ..., p(\mathbf{O}, Y||M_n)$, a manipulation M, background assumptions, and a target variable Y whose post-manipulation value is to be predicted.

Output: $\hat{Y}(\mathbf{O}||M)$, a predictor of the value of Y from $\mathbf{O}$ after manipulation M that has a small MSPE.

[2] In order to avoid technicalities, we will assume that the set of values conditioned on do not have measure 0.

[3] Other measures of prediction error, such as the absolute value of prediction error or optimizing certain decision problems could be used, but would not substantially change the general approach taken here.

[4] Here, p' is not a derivative of p; the prime after the p merely indicates that a new function has been introduced. The use of manipulated probability densities in decision theory is often not explicit. The assumption that the density of states of nature are independent of the actions taken (act-state independence) is one way to ensure that the manipulated densities that are needed are equal to observed conditional densities that can be measured.

Problem 2a: Constructing Causal Models from Sample Data

Input: i.i.d. samples from a population with density $p(\mathbf{O})$, a (possibly empty) set of i.i.d. samples from manipulated densities $p(\mathbf{O}||M_1), ..., p(\mathbf{O}||M_n)$, and background assumptions.

Output: A set of causal models that is as small as possible, and contains an approximately true causal model.

Problem 2b: Predicting the Effects of Manipulations from Causal Models

A set $\mathbf{C}$ of causal models over a set of variables $\mathbf{O}$ and Y, a manipulation M, and a target variable Y.

Output: $\hat{Y}(\mathbf{O}||M)$ if one exists, and an output of "no function" otherwise.

The reason that the stated goal for the output of Problem 2a is a set of causal models, rather than a single causal model is that it is in some cases it is not possible to reliably find a true causal model given the inputs. Furthermore, in contrast to predictive models, even if it a true causal model can be inferred from a sample from the unmanipulated population, it generally cannot be validated on a sample from the unmanipulated population, because a causal model contains predictions about a manipulated population that might not actually exist. This has been a serious impediment to the improvement of algorithms for constructing causal models, because it makes evaluating the performance of such algorithms difficult. It is possible to evaluate causal inference algorithms on simulated data, to employ background knowledge to check the performance of algorithms, and to conduct limited (due to expense, time, and ethical constraints) experiments, but these serve as only partial checks how algorithms perform on real data in a wide variety of domains.

3. STRUCTURAL EQUATION MODELS

The set of random variables in a structural equation model (SEM) can be divided into two subsets, the "error variables" or "error terms," and the substantive variables (for which there is no standard terminology in the literature). The substantive variables are the variables of interest, but they are not necessarily all observed. Each substantive variable X is a function of other substantive variables $\mathbf{V}$, and a unique error term ε_X; i.e., $X := f(\mathbf{V}, \varepsilon_X)$. We use an assignment operator, rather than an equality operator because the equations are interpreted causally; manipulating a variable in $\mathbf{V}$ can lead to a change in the value of X.

Each SEM is associated with a directed graph whose vertices include the substantive variables, and that represents both the causal structure of the model and the form of the structural equations. There is a directed edge from A to B ($A \rightarrow B$) if the coefficient of A in the structural equation for B is non-zero. In a linear SEM, the coefficient $b_{B,A}$ of A in the structural equation for B is the *structural coefficient associated* with the edge $A \rightarrow B$. In general, the graph of a SEM may have cycles (i.e., directed paths from a variable to itself), and may explicitly include error terms with double-headed arrows between them to represent that the error terms are dependent (e.g., $\varepsilon_A \leftrightarrow \varepsilon_B$); if no such edge exists in the graph, the error terms are assumed to be independent. If a variable has no arrow directed into it, then it is *exogenous*; otherwise it is *endogenous*. In SEM $K(\boldsymbol{\theta})$ depicted in Figure 1(i) (where $\boldsymbol{\theta}$ is the set of parameter values for K) A is exogenous and B and R are endogenous. If the graph has no directed cycles and no double-headed arrows, then it is a *directed acyclic graph* (DAG).

Given independent error terms in SEM K, for each $\boldsymbol{\theta}$, SEM K entails both a set of conditional independence relations among the substantive variables, and that the joint density over the substantive variables *factors according to the graph*, i.e., the joint density can be expressed as the product of the density of each variable conditional on its parents in the graph. For example, $p(A, B, R) = p(A)p(B|A)p(R|A)$ for all $\boldsymbol{\theta}$. This factorization in turn is equivalent to a set of conditional independence relations among the substantive variables [21].

$I_p(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ denotes that $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ in density p, i.e., $p(\mathbf{X}|\mathbf{Y}, \mathbf{Z}) = p(\mathbf{X}|\mathbf{Z})$ for all $p(\mathbf{X}|\mathbf{Z}) \neq 0$. (In cases where it does not create any ambiguity, the subscript p will be dropped). If a SEM M with parameter values $\boldsymbol{\theta}$ (represented by $M(\boldsymbol{\theta})$) entails that $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$, we write $I_{M(\boldsymbol{\theta})}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$. If a SEM with fixed causal graph M entails that $I_{M(\boldsymbol{\Theta})}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ for all possible parameter values $\boldsymbol{\Theta}$ we write $I_M(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$. In that case we say that M entails $I(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$. It is possible to determine whether $I_M(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ from the graph of M using the purely graphical criterion, "d-separation" [26].

A Bayesian network is a pair $\langle G, p\rangle$, where G is a directed acyclic graph and a p is a probability density such that if $\mathbf{X}$ and $\mathbf{Y}$ are d-separated conditional on $\mathbf{Z}$ in G, then $\mathbf{X}$ and $\mathbf{Y}$ are independent conditional on $\mathbf{Z}$ in G. If the error terms in a SEM with a DAG G are jointly independent, and $p(\mathbf{V})$ is the entailed density over the substantive variables, then $\langle G, p(\mathbf{V})\rangle$ is a Bayesian network.

3.1 Representing Manipulations in a SEM

Given a linear SEM, a manipulation of a variable X_i in a population can be described by the following kind of equation: $X_i = \sum_{X_j \in \mathbf{PA}(X_i)} b_{i,j} X_j + \varepsilon_i$, where all of the variables are the post-manipulation variables, $\mathbf{PA}(X_i)$ is a new set of causes of X_i (which are included in the set of non-effects of X_i in the unmanipulated population). A simple special case is where X_i is set to a constant c.

In a causal model such as SEM $K(\boldsymbol{\theta})$, the post-manipulation population is represented in the following way, as shown in Figure 1. The result of modifying the set of structural equations in this way can lead to a density in the randomized population that is not necessarily the same as the density in any subpopulation of the general population. (For more details see [27, 36].) See Figure 1 for examples of manipulations to SEM K.

A set $\mathbf{S}$ of variables is *causally sufficient* if every variable H that is a direct cause (relative to $\mathbf{S} \cup \{H\}$) of any pair of variables in $\mathbf{S}$ is also in $\mathbf{S}$. Intuitively, a set of variables $\mathbf{S}$ is causally sufficient if no common direct causes (relative to $\mathbf{S}$) have been left out of $\mathbf{S}$. If SEM K is true then $\{A, B, R\}$ is causally sufficient, but $\{B, R\}$ is not because A is a common direct cause of B and R relative to $\{A, B, R\}$ but is not in $\{B, R\}$. If the observed set of variables is not causally sufficient, then the causal model is said to contain *unobserved common causes*, *hidden common causes*, or *latent variables*.

4. ASSUMPTIONS

The following assumptions are often used to relate causal

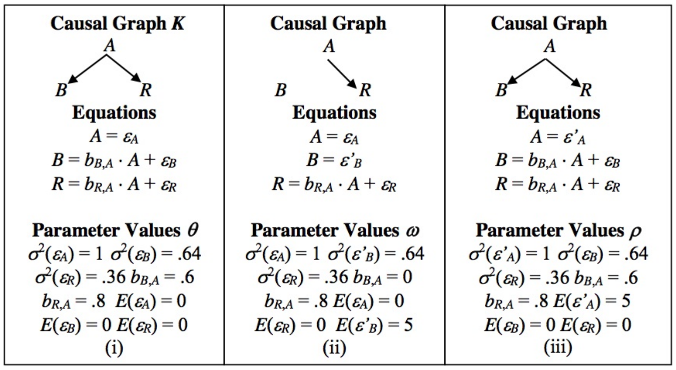

Causal Graph K
A
B
R
Equations
$A = \varepsilon_A$
$B = b_{B,A} \cdot A + \varepsilon_B$
$R = b_{R,A} \cdot A + \varepsilon_R$
Parameter Values θ
$\sigma^2(\varepsilon_A) = 1$ $\sigma^2(\varepsilon_B) = .64$
$\sigma^2(\varepsilon_R) = .36$ $b_{B,A} = .6$
$b_{R,A} = .8$ $E(\varepsilon_A) = 0$
$E(\varepsilon_B) = 0$ $E(\varepsilon_R) = 0$
(i)
Causal Graph
A
B
R
Equations
$A = \varepsilon_A$
$B = \varepsilon'_B$
$R = b_{R,A} \cdot A + \varepsilon_R$
Parameter Values ω
$\sigma^2(\varepsilon_A) = 1$ $\sigma^2(\varepsilon'_B) = .64$
$\sigma^2(\varepsilon_R) = .36$ $b_{B,A} = 0$
$b_{R,A} = .8$ $E(\varepsilon_A) = 0$
$E(\varepsilon_R) = 0$ $E(\varepsilon'_B) = 5$
(ii)
Causal Graph
A
B
R
Equations
$A = \varepsilon'_A$
$B = b_{B,A} \cdot A + \varepsilon_B$
$R = b_{R,A} \cdot A + \varepsilon_R$
Parameter Values ρ
$\sigma^2(\varepsilon'_A) = 1$ $\sigma^2(\varepsilon_B) = .64$
$\sigma^2(\varepsilon_R) = .36$ $b_{B,A} = .6$
$b_{R,A} = .8$ $E(\varepsilon'_A) = 5$
$E(\varepsilon_B) = 0$ $E(\varepsilon_R) = 0$
(iii)

Figure 1: (i) Unmanipulated causal graph K; (ii) B Manipulated to 5; (iii) A Manipulated to 5

relations to probability densities.

4.1 The Causal Markov Assumption

Causal Markov Assumption: For causally sufficient sets of variables, all variables are independent of the their non-effects (non-descendants in the causal graph) conditional on their direct causes (parents in the causal graph) [36].

The Causal Markov Assumption is an oversimplification because it basically assumes that all associations between variables are due to causal relations. There are several other ways that associations can be produced.

First, conditioning on a common descendant can produce a conditional dependency. For example, if sex and intelligence are unassociated in the population, but only the most intelligent women attend graduate school, while men with a wider range of intelligence attend graduate school, then sex and intelligence will be associated in a sample consisting of graduate students (i.e., sex and intelligence cause graduate school attendance, which has been conditioned on in the sample.) See [37] for a discussion of selection bias. Second, logical relationships between variables can also produce non-causal correlations (e.g., if GDP_yearly is defined to be the sum of $GDP\_January$, $GDP\_Februrary$, etc., GDP_yearly will be associated with these variables, but not caused by them.) For a discussion of logical relations between variables, see [38]. Third, it does not have any way of dealing with instantaneous symmetric interactions (like classical theories of gravity).

4.2 The Causal Faithfulness Assumption

Consider SEM O in Figure 2. Suppose we have $I_K(B, R|A)$, where SEM K is shown in Figure 1(i), whereas it is not the case that $I_O(B, R|A)$. However, just because O does not entail $I_O(B, R|A)$ for all sets of parameter values $\boldsymbol{\beta}$, that does not imply that there are *no* $\boldsymbol{\beta}$ for which $I_{O(\boldsymbol{\beta})})(B, R|A)$. For example, if the variances of R, A, and B are all 1, for any $\boldsymbol{\beta}$ for which $\text{cov}_{O(\boldsymbol{\beta})}(A, B) \cdot \text{cov}_{O(\boldsymbol{\beta})}(A, R) = \text{cov}_{O(\boldsymbol{\beta})}(B, R)$, it follows that $\text{cov}_{O(\boldsymbol{\beta})}(B, R|A) = 0$. This occurs when $(b_{B,R} \cdot b_{A,R} + b_{A,B}) \cdot (b_{B,R} \cdot b_{A,B} + b_{A,R}) = b_{R,B}$. So if $I_p(B, R|A)$ is true in the population, there are at least two kinds of explanation: any set of parameter values for SEMs K (in Figure 1(i)), L, or M (in Figure 2) on the one hand, or any parameterization of SEM O for which $(b_{B,R} \cdot b_{A,R} + b_{A,B}) \cdot (b_{B,R} \cdot b_{A,B} + b_{A,R}) = b_{R,B}$. There are several arguments why, although O with the special parameter values is a possible explanation, in the absence of evidence to the contrary, K, L, or M should be the preferred explanations.

First, K, L, and M explain the independence of B and R conditional on A structurally, as a consequence of no direct causal connection between the variables. In contrast O explains the independence as a consequence of a large direct effect of B on R cancelled *exactly* by the product of large direct and indirect effects of B and R on A.

Second, this cancellation is improbable (in the Bayesian sense that if a zero conditional covariance is not entailed, the measure of the set of free parameter values for any DAG that lead to such cancellations is zero for any "smooth" prior probability density,[5] such as the Gaussian or exponential one, over the free parameters).

Finally, K, L, and M are simpler than O. K, L, and M have fewer free parameters than O.

The assumption that a causal influence is not hidden by coincidental cancellations can be expressed for SEMs in the following way. A density p is *faithful* to the graph G of a SEM if and only if every conditional independence relation true in p is entailed by G.

Causal Faithfulness Assumption: For a causally sufficient set of variables $\mathbf{V}$ in a population P, the population density $p_P(\mathbf{V})$ is faithful to the causal graph over $\mathbf{V}$ for P [36].

The Causal Faithfulness Assumption requires preferring K, L, and M to O, because parameter values $\boldsymbol{\beta}$ for which $I_{O(\boldsymbol{\beta})})(B, R|A)$ would violate the Causal Faithfulness Assumption. Recently, there have been a number of search algorithms that are consistent, but have substituted other kinds of assumptions in place of the Causal Faithfulness Assumption.

4.3 The Output of a Search for Causal Models

The following sections describe several different possible alternatives that can be output by a reliable search algorithm.

4.3.1 Markov Equivalence Classes

A trek between A and B is either a directed path from A to B, a directed path from B to A, or a path between A and B that does not contain a subpath $X \to Y \leftarrow Z$. SEMs K, L, and M are Markov equivalent, in the sense that their respective graphs all entail the same set of conditional independence relations. If K is true, any SEM with a graph that contains no path between A and R can be eliminated from consideration by the Causal Markov Assumption (e.g., N in Figure 2). SEM P also violates the Causal Markov Assumption. O is incompatible with the population conditional independencies by the Causal Faithfulness Assumption. However, neither of these assumptions implies L or M is incompatible with the population conditional independencies.

Since K, L, and M entail the same set of conditional independence relations, it is not possible to eliminate L or M as incompatible with the population conditional independence relations without either adding more assumptions or background knowledge, or using features of the probability density that are not conditional independence relations. In the case of linear SEMs with Gaussian error terms (and for multinomial Bayesian networks) there are no other features

[5] A smooth measure is absolutely continuous with Lebesgue measure.

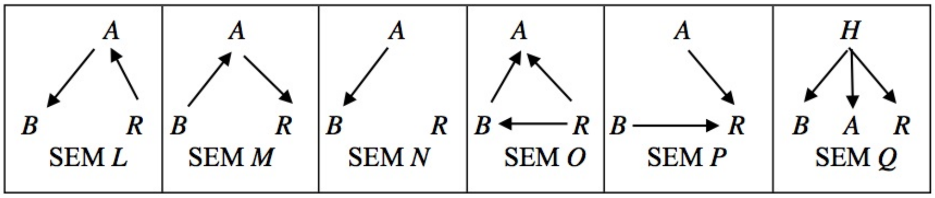


Figure 2: Alternative SEM models

of the density that distinguish K from L or M. However, as we will illustrate later, for other families of distributions, there are non-conditional independence constraints that can be entailed by a graph that do distinguish K from L or M.

4.3.2 Distribution Equivalence

K and L are *distribution equivalent* if and only if for any assignment of parameter values $\boldsymbol{\theta}$ to K there exists an assignment of parameter values $\boldsymbol{\theta}'$ to L that represents the same density, and vice versa. If all of the error terms are Gaussian with linear causal relations, then K and L are distribution equivalent as well as Markov equivalent. In such cases, the best that a reliable search algorithm can do is to return the entire Markov equivalence class, regardless of what features of the marginal density that it uses.

In contrast, for linear causal models with at most one error term is non-Gaussian, SEMs K and L are Markov equivalent, but they are not distribution equivalent.

When Markov equivalence fails to entail distribution equivalence, then using conditional independence relations alone for causal inference is still correct, but it is not as informative as theoretically possible. For example, assuming linearity, causal sufficiency, and non-Gaussian errors [32], conditional independence tests can at best reliably determine the correct Markov equivalence class, while using other features of the sample density can be used to reliably determine a unique graph [32].

4.4 Constraint-Based Search

The number of DAGs grows super-exponentially with the number of vertices, so even for modest numbers of variables it is not possible to examine each DAG to determine whether it is compatible with the population density given the Causal Markov and Faithfulness Assumptions. The PC algorithm, given as input an oracle that returns answers about conditional independence in the population and optional background knowledge about orientations of edges, returns a graphical object called a pattern that represents a Markov equivalence class (or if there is background knowledge a subset of a Markov equivalence class) on the basis of oracle queries. If the oracle always gives correct answers, and the Causal Markov and Causal Faithfulness Assumptions hold, then the output pattern contains the true SEM, even thought the algorithm does not check each DAG. In the worse case, it is exponential in the number of variables, but for sparse graphs it can run on hundreds of thousands of variables [34, 35, 23].

5. DIFFERENCES BETWEEN CLASSIFICATION AND REGRESSION AND CAUSAL INFERENCE

The following is a brief summary of some important differences between the problem of predicting the value of an variable in an unmanipulated population from a sample, and the problem of predicting the post-manipulation value of a variable from a sample from an unmanipulated population. In an unmanipulated population P, the predictor that minimizes the MSPE is the conditional expected value.

1. $E(Y|\mathbf{O})$ (the expected value of Y conditional on $\mathbf{O}$) is a function of $p(\mathbf{O},Y)$, regardless of what the true causal model is.[6] In contrast, a manipulated expected value is a function of $p(\mathbf{O}, Y)$ *and* a causal graph.

2. In order to determine whether $E_P(Y||p'(\mathbf{O}))$ (the expected value of Y after a manipulation to $p'(\mathbf{O})$) is a function of $p(\mathbf{O}, Y)$ and background knowledge, it is necessary to find *all* of the causal models compatible with $p(\mathbf{O}, Y)$ and background knowledge, not simply one causal model compatible with $p(\mathbf{O}, Y)$ and background knowledge.

3. Determining which causal models are compatible with background knowledge and a $p(\mathbf{O}, Y)$ requires making additional assumptions connecting population densities to causal models (e.g., Causal Markov and Faithfulness).

4. Without introducing some simplicity assumptions about causal models, for some common families of densities (e.g., Gaussian, multinomial), no $E_P(Y|\mathbf{O}'||p'(O))$ are functions of the population density without very strong background knowledge.

5. The justification for using simple statistical models is fundamentally different than the justification for using simple causal models. At a given sample size, the use of simple statistical model can be justified even if causal relations are not simple. However, the assumption that the simplest causal model compatible with $p(\mathbf{O}, Y)$ and background knowledge is a substantive assumption about the simplicity of mechanisms that exist in the world.

6. For many families of densities (e.g., Gaussian, multinomial), there is always a statistical model without hidden variables that contains the population density. For those same families of densities, a causal model that contains both the population probability density and the post-manipulation probability densities may require the introduction of unobserved variables.

[6] This ignores the problem of conditioning on sets of measure zero.

7. Given a population density, and the set of causal models consistent with the population density and background knowledge, calculating the effects of a manipulation can be difficult because:

    a. There may be unobserved variables (even if only a single causal model is consistent with $p(\mathbf{O}, Y)$ and background knowledge).

    b. There may be multiple causal models compatible with $p(\mathbf{O}, Y)$ and background knowledge.

8. For non-experimental data, a post-manipulation density is different than the population density from which the sample is drawn. The post-manipulation values of the target variable Y are not directly measured in the sample. Hence, it is not possible to estimate the error in $E_P(Y|\mathbf{O}'||p'(O))$ by comparing it to the values in a sample from the $p(\mathbf{O}, Y)$.

6. SEMS CAN HELP IN CAUSAL DISCOVERY FROM I.I.D. AND TIME SERIES DATA

As discussed in Section 4.4, the constraint-based approach to causal discovery involves conditional independence tests, which would be a difficult task if the form of dependence is unknown. It has the advantage that it is generally applicable, but the disadvantages are that faithfulness is a strong assumption and that it may require very large sample sizes to get good conditional independence tests. Furthermore, the solution of this approach to causal discovery is usually non-unique, and in particular, it does not help on determining causal direction in the two-variable case, where no conditional independence relationship is available.

What information can we use to fully determine the causal structure? A fundamental issue is given two variables, how to distinguish cause from effect. To do so, one needs to find a way to capture the asymmetry between them. Intuitively, one may think that the physical process that generates effect from cause is more natural or simple in some way than recovering the cause from effect. How can we represent this generating process, and in which way is the causal process more natural or simple than the backwards process?

Recently several causal discovery approaches based on structural equation models (SEMs) have been proposed. A SEM represents the effect Y as a function of the direct causes X and some unmeasurable error:

$$Y = f(X, \varepsilon; \boldsymbol{\theta}_1), \tag{1}$$

where ε is the error term that is assumed to be independent from X, the function $f \in \mathcal{F}$ explains how Y is generated from X, $\mathcal{F}$ is an appropriately constrained functional class, and $\boldsymbol{\theta}_1$ is the parameter set involved in f. We assume that the transformation from (X, ε) to (X, Y) is invertible, such that N can be uniquely recovered from the observed variables X and Y.

For convenience of presentation, let us assume that both X and Y are one-dimensional variables. Without precise knowledge on the data-generating process, the SEM should be flexible enough such that it could be adapted to approximate the true data-generating process; more importantly, the causal direction implied by the SEM has to be identifiable in most cases, i.e., the model assumption, especially the independence between the error and cause, holds for only one direction, such that it implies the causal asymmetry between X and Y. Under the above conditions, one can then use SEMs to determine the causal direction between two variables, given that they have a direct causal relationship in between and do not have any confounder: for both directions, we fit the SEM, and then test for independence between the estimated error term and the hypothetical cause, and the direction which gives an independent error term is considered plausible.

Several forms of the SEM have been shown to be able to produce unique causal directions, and have received practical applications. In the linear, non-Gaussian, and acyclic model (LiNGAM [32]), f is linear, and at most one of the error term ε and cause X is Gaussian. The nonlinear additive noise model [15, 44] assumes that f is nonlinear with additive noise (error) ε. In the post-nonlinear (PNL) causal model [45], the effect Y is further generated by a post-nonlinear transformation on the nonlinear effect of the cause X plus error term ε:

$$Y = f_2(f_1(X) + \varepsilon), \tag{2}$$

where both f_1 and f_2 are nonlinear functions and f_2 is assumed to be invertible.[7] The post-nonlinear transformation f_2 represents the sensor or measurement distortion, which is frequently encountered in practice. In particular, the PNL causal model has a very general form (the former two are its special cases), but it has been shown to be identifiable in the generic case (except five specific situations given in [45]). It is worth noting that it is not closed under marginalization, even if there are not confounders. In the subsequent sections, we will discuss the identifiability of various SEMs, how to distinguish cause from effect with the SEMs, and the relationships between different principles for causal discovery, including mutual independence of the error terms and the causal Markov condition, respectively.

Another issue we are concerned with is causal discovery from time series. According to [13], Granger's causality in time series falls into the framework of constraint-based causal discovery combined with the temporal constraint that the effect cannot precede the cause. The SEM, together with the above temporal constraint, has also been exploited to estimate time-delayed causal relations possibly with instantaneous effects [44]. Compared to the conditional independence relationships, the SEM, if correctly specified, is able to recover more about the causal information. In this paper, when talking about causality in time series, we assume that the causal relations are linear with non-Gaussian errors. In Section 10, after reviewing linear Granger causality with instantaneous effects, we focus on two problems which are traditionally difficult to solve. In particular, we present the theoretical results which make it possible to discover the temporal causal relations at the true causal frequency from subsampled data [12], that is, one can recover monthly causal relations from quarterly data or estimate rapid causal influences between stocks from their daily returns. Moreover, even when there exist confounder time series, theoretical results suggested that one can still identify the causal

[7] In [42] both functions f_1 and f_2 are assumed to to invertible; this causal model, as a consequence, can be estimated by making use of post-nonlinear independent component analysis (PNL-ICA) [39], which assumes that the observed data are component-wise invertible transformations of linear mixtures of the independence sources to be recovered.

relations among the observed time series as well as the influences from the confounder series [11].

7. SEVERAL SEMS AND THE IDENTIFIABILITY OF CAUSAL DIRECTION

When talking about the causal relation between two variables, traditionally people were often concerned with the linear-Gaussian case, where the involved variables are Gaussian with a linear causal relation, or the discrete case. It turned out that the former case is one of the atypical situations where the causal asymmetry does not leave a footprint in the observed data or their joint distribution: the joint Gaussian distribution is fully determined by the mean and covariance, and with proper rescaling, the two variables are completely asymmetric w.r.t. the data distribution.

In the discrete case, if one knows precisely what SEM class generated the effect from cause, which, for instance, may be the noisy AND or noisy XOR gate, then under mild conditions the causal direction can be easily seen from the data distribution. However, generally speaking, if the precise functional class of the causal process is unknown, in the discrete case it is difficult to recover the causal direction from observed data, especially when the cardinality of the variables is small. As an illustration, let us consider the situation where the causal process first generates continuous data and discretizes such data to produce the observed discrete ones. It is then not surprising that certain properties of the causal process are lost due to discretization, making causal discovery more difficult. In this paper we will focus on the continuous case.

7.1 Causal Direction Is Not Identifiable without Constraints on SEMs

In the SEM (1), the error term is assumed to be independent from the cause. If for the reverse direction, one cannot find a function to represent X in terms of the hypothetical cause Y and an error term which is independent from Y, then we can determine the true causal direction, or distinguish cause from effect. Unfortunately, this is not the case if we do not impose any constraint on the function f, as explained below.

According to [17], given *any* two random variables X and Y with continuous support, one can always construct another variable, denoted by $\tilde{\varepsilon}$, which is statically independent from X. In [47] the class of functions to produce such an independent variable $\tilde{\varepsilon}$ (or called independent error term in our causal discovery context) was given, and it was shown that this procedure is invertible: Y is a function of X and $\tilde{\varepsilon}$.

This is also the case for the hypothetical causal direction $Y \to X$: we can also always represent X as a function of Y and an independent error term. That is, any two variables would be symmetric according to the SEM, if f is not constrained. Therefore, in order for the SEMs to be useful to determine the causal direction, we have to introduce certain constraints on the function f such that the independence condition on the error and the hypothetical cause holds for only one direction. Below we focus on the two-variable case, and the results can be readily extended to the case with an arbitrary number of variables, as shown in [28].

7.2 Linear Non-Gaussian Causal Model

The linear causal model in the two-variable case can be written as

$$Y = bX + \varepsilon. \tag{3}$$

It is nice that if at most one of X and ε is Gaussian, the causal direction is identifiable, due to the ICA theory [16], or more fundamentally, due to the Darmois-Skitovich theorem [20]. This is known as the linear, non-Gaussian, acyclic model (LiNGAM [32]).

7.2.1 *On the Ubiquitousness of Non-Gaussianity in the Linear Case*

According to the central limit theorem, under mild conditions, the sum of independent variables tends to be Gaussian as the number of components becomes larger and larger. One may then challenge the non-Gaussianity assumption in the LiNGAM model. Here we argue that in the linear case, non-Gaussian distributions are ubiquitous.

Cramér's decomposition theorem states that if the sum of two independent real-valued random variables is Gaussian, then both of the summand variables much be Gaussian as well; see [6, page 53]. By induction, this means that if the sum of any finite independent real-valued variables is Gaussian, then all summands must be Gaussian. In other words, a Gaussian distribution can never be exactly produced by linear composition of variables any of which is non-Gaussian. This nicely complements the central limit theorem: (under proper conditions) the sum of independent variable get closer to Gaussian, but it cannot be exactly Gaussian, except all summand variables are Gaussian. This linear closure property of the Gaussian distribution implies the rareness of the Gaussian distribution and ubiquitousness of non-Gaussian distributions, if we believe the relations between variables are linear. However, the closer it gets to Gaussian, the harder it is to distinguish the direction. Hence, the practical question is, are the errors typically non-Gaussian enough to distinguish causal directions in the linear case?

7.3 Nonlinear Additive Noise Model

In practice nonlinear transformation is often involved in the data generating process, and should be taken into account in the functional class. As a direct extension of LiNGAM, the nonlinear additive noise model represents the effect as a nonlinear function of the cause plus independent error [15]:

$$Y = f_{AN}(X) + \varepsilon. \tag{4}$$

It has been shown that the set of all $p(X)$ for which the backward model also admits an independent error term is contained in a 3-dimensional affine space. Bearing in mind that the space of all possible $p(X)$ is infinite dimensional, one can see that roughly speaking, in the generic case, if the data were generated by the nonlinear additive noise model, the causal direction is identifiable. This model is a special case of the PNL causal model, which is to be discussed below, and the identifiability results for the PNL causal model also apply here.

7.4 Post-Nonlinear Causal Model

If the assumed SEM is too restrictive to be able to approximate the true data generating process, the causal discovery results may be misleading. Therefore, if the specific knowledge about the data generating mechanism is not available, to make it useful in practice, the assumed causal model should be general enough, such that it can reveal the data generating processes approximately.

The PNL causal model takes into account the nonlinear influence from the cause, the noise effect, and the possible sensor or measurement distortion in the observed variables [45]. See (2) for its form; a slightly more restricted version of the model, in which the inner function, f_1, is also assumed to be invertible, was proposed in [42] and applied to causal analysis of stock returns. It has the most general form among all well-defined SEMs according to which the causal direction is identifiable in the general case. (The model used in [24] does not impose structural constraints but assumes a certain type of smoothness; however, it does not lead to theoretical identifiability results.) Clearly it contains the linear model and nonlinear additive noise model as special cases. The multiplicative noise model, $Y = X \cdot \varepsilon$, where all involved variables are positive, is another special case, since it can be written as $Y = \exp(\log X + \log \varepsilon)$, where $\log \varepsilon$ is considered as a new noise term, $f_1(X) = \log(X)$, and $f_2(\cdot) = \exp(\cdot)$.

The identifiability conditions of the causal direction according to the PNL causal model was established by a proof by contradiction [45]. We assume the causal model holds in both directions $X \to Y$ and $Y \to X$, and show that this implies some very strong conditions on the distributions and functions involved in the model. Suppose the data were generated according to the PNL causal model in settings other than those specific conditions; then in principle, the backward direction does not follow the model, and the causal direction can be determined.

Assume that the data (X, Y) are generated by the PNL causal model with the the causal relation $X \to Y$. This data generating process can be described as (2). Moreover, let us assume that the backwards direction, $Y \to X$ also follows the PNL causal model with independent error. That is,

$$X = g_2(g_1(Y) + \varepsilon_Y), \qquad (5)$$

where Y and ε_Y are independent, g_1 is non-constant, and g_2 is invertible.

Equations (2) and (5) define the transformation from $(X, \varepsilon)^\intercal$ to $(Y, \varepsilon_Y)^\intercal$; as a consequence, $p(Y, \varepsilon_Y)$ can be expressed in terms of $p(X, \varepsilon) = p(X)p(\varepsilon)$. The identifiability results were obtained based on the linear separability of the logarithm of the joint density of independent variables, i.e., for a set of independent random variables whose joint density is twice differentiable, the Hessian of the logarithm of their density is diagonal everywhere [22]. Since Y and ε_Y are assumed to be independent, $\log p(Y, \varepsilon_Y)$ then follows such a linear separability property. This implies that the second-order partial derivative of $\log p(Y, \varepsilon_Y)$ w.r.t. Y and ε_Y is zero. It then reduces to a differential equation of a bilinear form. Under certain conditions (e.g., $p(\varepsilon)$ is positive on $(-\infty, +\infty)$), the solution to the differential equation gives all cases in which the causal direction is *not* identifiable according to the PNL causal model. Table 1 in [45] summarizes all five non-identifiable cases. The first one is the linear-Gaussian case, in which the causal direction is well-known to be non-identifiable. Roughly speaking, to make one of those cases true, one has to adjust the data distribution and the involved nonlinear functions very carefully. In other words, in the generic case the causal direction is identifiable if the data were generated according to the PNL causal model.

8. DETERMINATION OF CAUSAL DIRECTION BASED ON SEMS

A commonly used approach to distinguishing cause from effect with nonlinear SEMs consists of two steps. First, one fits the model (e.g., the nonlinear additive noise model or the PNL causal model) on the data for both hypothetical causal directions. The second step is to do independence test between the estimated error term and hypothetical cause [15, 45]. If the independence condition holds for one and only one hypothetical direction, the causal relation between the two variables X and Y implied by the corresponding SEM has been successfully found. If neither of them holds, the data-generating process may not follow the assumed SEM, or there exists some confounder influencing both X and Y. If both hold, the cause and effect can not be distinguished by the exploited SEM; in this case, additional information, such as the smoothness of the involved nonlinearities, may help find the causal model with a lower complexity. We adopted the Hilbert Schmidt information criterion (HSIC) [14] for statistical independence test in the first step. Below we discuss how to estimate the function as well as the error term in the first step.

For the nonlinear additive noise model, the function f_{AN} is usually estimated by performing Gaussian process (GP) regression [15]. For details on GP regression, one may refer to [29].

Estimation of the PNL causal model (2) has several indeterminacies: the sign, mean, and scale of the error term $varepsilon$, and accordingly, the sign, location, and scale of f_{i1} are arbitrary. In the estimation procedure, one may impose certain constraints to avoid such indeterminacies in the estimate. However, we should note that in principle, we do not care about those indeterminacies in the causal discovery context, since they do not change the statistical independence or dependence property between the estimated error term and the hypothetical cause.

It is well known that for linear regression, the maximum likelihood estimator of the coefficient is still statistically consistent even if the error distribution is wrongly assumed to the Gaussian. However, this may not be the case for general nonlinear models. As shown in [47, Section 3.2], if the error distribution mis-specified, the estimated PNL causal model (2) may not be statistically consistent, even when the above indeterminacies in the estimate are properly tackled. Therefore, the error distribution should be adaptively estimated from data, if the true one is not known *a priori*. It has been proposed to estimate the PNL causal model (2) by mutual information minimization [45] with the involved nonlinear functions represented by multi-layer perceptrons (MLPs). Later, in [47] the PNL causal model was estimated by extending the framework of warped Gaussian processes to allow a flexible error distribution, which is represented by a mixture of Gaussians (MoG).

9. ON THE RELATIONSHIPS BETWEEN DIFFERENT PRINCIPLES FOR MODEL ESTIMATION

One usually use maximum likelihood to fit the SEM together with a directed acyclic graph (DAG) to the given data. Not surprisingly, the negative likelihood (with the distribution of the error term adaptively estimated from data) is equivalent to the mutual information between the estimated error terms, as stated in Theorem 3 in [47]. The higher the likelihood, the less dependent the estimated error terms. (Note that the root variables in the DAG are also counted as error terms.)

On the other hand, the constraint-based approach to causal discovery exploits conditional independence relationships of the variables to derive (the equivalence class of) the causal structure [36, 27]. How are these principles, including mutual independence of the estimated error terms and the causal Markov condition, related to each other? Below we will answer this question, and the results in this section hold for an arbitrary number of variables.

Let us consider optimization over different DAG structures to find the causal structure. Assume the we optimally fit the nonlinear functions f_i according to the given candidate DAG structure. First consider the situation where we fit the nonlinear additive noise model, i.e.,

$$X_i = f_{AN,i}(\mathbf{PA}_i) + \varepsilon_i, \tag{6}$$

to the data. It has been shown that mutual independence of the error terms and conditional independence between observed variables (together with the independence between ε_i and $\mathbf{PA}_i$) are equivalent. Furthermore, they are achieved if and only if the total entropy of the disturbances is minimized [44]. More specifically, when fitting the model (6) with a hypothetical DAG causal structure to the given variables $X_1, \cdots, X_n$, the following three properties are equivalent:

- (i) The causal Markov condition holds (i.e., each variable is independent of its non-descendants in the DAG conditioning on its parents), and in addition, the error term in X_i is independent from the parents of X_i.
- (ii) The error terms N_i are mutually independent.
- (iii) The total entropy of the error terms, i.e., $\sum_i H(\varepsilon_i)$, is minimized, with the minimum $H(X_1, \cdots, X_n)$.

Let us then consider the PNL causal model. When one fits the PNL causal model

$$X_i = f_{i2}(f_{i1}(\mathbf{PA}_i) + \varepsilon_i), \tag{7}$$

to the data, the scale of the error terms as well as f_{i1} is arbitrary, since f_{i2} is also to be estimated. Consequently, unlike for the nonlinear additive noise model, in the PNL causal model context it is not meaningful to talk about the total entropy of the error terms (see condition (iii) above). However, as shown in [45], when fitting the PNL causal model with a hypothetical DAG causal structure to the data, we still have the equivalence between conditions (i) and (ii) above.

Given more than two variables, one way to estimate the causal model based on SEMs is to use exhaustive search: for all possible causal orderings, fit SEMs for all hypothetical effects separately, and then do model checking by testing for independence between the estimated error and the corresponding hypothetical causes. However, note that the complexity of this procedure increases super-exponentially along with the number of variables. Smart approaches are then needed.

The above result concerning the relationship between mutual independence of the error terms and the causal Markov condition combined with the independence between each error term and its associated parents suggests a two-step method to find the causal structure implied by the PNL causal model. One first uses the constraint-based approach to find the Markov equivalent class from conditional independence relationships with proper nonparametric conditional independence tests (e.g., [46]). The PNL causal model is then used to identify the causal directions that cannot be determined in the first step: for each DAG contained in the equivalent class, we estimate the error terms, and determine whether this causal structure is plausible by examining whether the disturbance in each variable X_i is independent from the parents of X_i. Consequently, one avoids the exhaustive search over all possible causal structures and high-dimensional statistical tests of mutual independence of all error terms.

10. CAUSAL DISCOVERY FROM TIME SERIES

Both the constraint-based and SEM-based approaches to causal discovery are directly applicable to find causal relations from time series; moreover, one can benefit from the temporal constraint that the effect cannot precede the cause, which helps reduce the search space of the causal structure. Below we assume linearity of the causal relations and consider three problems, namely, linear Granger causal analysis with instantaneous effects, causal discovery from systematically subsampled data, and that in the presence of hidden time series.

10.1 Linear Granger Causality and its Extension with Instantaneous Effects

For Granger causal analysis in the linear case [13], one fits the following VAR model [33] to the data:

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{8}$$

where $\mathbf{X}_t = (X_{1t}, X_{2t}, ..., X_{nt})^\intercal$ is the vector of the observed data, $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, ..., \varepsilon_{nt})^\intercal$ is the temporally and contemporaneously independent noise process, and causal transition matrix $\mathbf{A}$ contains the temporal causal relations.

In practice it is found that after fitting the VAR model, the residuals are often contemporaneously dependent. To account for such dependence, the above VAR model has been extended to allow instantaneous causal effects between X_{it} [18]. Let $\mathbf{B}_0$ contains the instantaneous causal relations between $\mathbf{X}_t$. Equation (8) changes to

$$\begin{aligned}
&\mathbf{X}_t = \mathbf{B}_0\mathbf{X}_t + \mathbf{A}\mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t,\\
\Rightarrow&(\mathbf{I} - \mathbf{B}_0)\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t,\\
\Rightarrow&\mathbf{X}_t = (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{A}\mathbf{X}_{t-1} + (\mathbf{I} - \mathbf{B}_0)^{-1}\boldsymbol{\varepsilon}_t.
\end{aligned} \tag{9}$$

To estimate all involved parameters in Granger causality with instantaneous effects, two estimation procedures have been proposed in [18]. The two-step method first estimate the errors in the above VAR model and then apply independent component analysis (ICA) [16] on the estimated errors. The other is based on multichannel blind deconvolution, which is statistically more efficient [44].

10.2 Causal Discovery from Subsampled Data

Suppose the original high-resolution data were generated by (8). We consider low-resolution data generated by subsampling (or systematic sampling) with the subsampling factor k. Here we are interested in finding the causal transition matrix $\mathbf{A}$ which generated the data from the subsampled data. Traditionally, if one uses only the second-order information, this suffers from parameter identification issues [25], i.e., the same subsampled (low-frequency) model may disaggregate to several high frequency models, which are observationally equivalent at the low frequency.

10.2.1 Effect of Subsampling (Systematic Sampling)

Suppose that due to low resolution of the data, there is an observation every k time steps. That is, the low-resolution observations $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, , ..., \tilde{\mathbf{X}}_t)$ are $(\mathbf{X}_1, \mathbf{X}_{1+k}, ..., \mathbf{X}_{1+(t-1)k})$; here we have assumed that the first sampled point is $\mathbf{X}\mathbf{x}_1$. We then have

$$\begin{aligned}\tilde{\mathbf{X}}_{t+1} &= \mathbf{X}_{1+tk} = \mathbf{A}\mathbf{X}_{1+tk-1} + \boldsymbol{\varepsilon}_{1+tk}\\ &= \mathbf{A}(\mathbf{A}\mathbf{X}_{1+tk-2} + \boldsymbol{\varepsilon}_{1+tk-1}) + \boldsymbol{\varepsilon}_{1+tk}\\ &= ...\\ &= \mathbf{A}^k\tilde{\mathbf{X}}_t + \underbrace{\sum_{l=0}^{k-1}\mathbf{A}^l\boldsymbol{\varepsilon}_{1+tk-l}}_{\triangleq\vec{\boldsymbol{\varepsilon}}_t}. \end{aligned} \tag{10}$$

According to (10), subsampled data $\tilde{\mathbf{X}}_t$ also follows a vector autoregression (VAR) model with the error term $\vec{\boldsymbol{\varepsilon}}_t$, and one can see that as $T \to \infty$, the discovered temporal causal relations from such subsampled data are given by $\mathbf{A}^k$. As $k \to \infty$, $\mathbf{A}^k$ tends to vanish, and the subsampled data will be contemporaneously dependent. (We have assumed that the system is stable, in that all eigenvalues of $\mathbf{A}$ have modulus smaller than one.)

Misleading Granger causal relations in low-resolution data: An illustration.

Suppose $A = \begin{bmatrix}0.8 & 0.5\\ 0 & -0.8\end{bmatrix}$. Consider the case where $k = 2$. The corresponding VAR model for the subsampled data is

$$\tilde{\mathbf{X}}_t = \mathbf{A}^2\tilde{\mathbf{X}}_{t-1} + \vec{\boldsymbol{\varepsilon}}_t = \begin{bmatrix}0.64 & 0\\ 0 & 0.64\end{bmatrix}\tilde{\mathbf{X}}_{t-1} + \vec{\boldsymbol{\varepsilon}}_t.$$

That is, the causal influence from $X_{2,t-1}$ to X_{1t} is missing in the corresponding low-resolution data (with $k = 2$).

10.2.2 Identifiability of the Causal Relations at the Causal Frequency

It has been shown that if the distributions p_{N_i} are non-Gaussian and different for different i, together with other technical assumptions, the transition matrix associated with the causal-frequency data, $\mathbf{A}$, is identifiable from the subsampled data $\tilde{\mathbf{X}}$. As a by-product, the result also indicates that the subsampled data, although contemporaneously dependent, actually *do not* follow the model of linear Granger causality with instantaneous effects [12].

Let the distributions of the noise terms be represented by the MoG. An EM algorithm and a variational EM (with mean field approximation) were then proposed to estimate $\mathbf{A}$ from subsampled data.

10.3 Causal Discovery with Hidden Time Series (Confounders)

In practice it is usually difficult and even impossible to collect all relevant time series when doing causal analysis on given ones. We approach this problem as follows. We assume that the (multivariate) measurements are a sample of a multivariate random process $\mathbf{X}_t$, which, together with another random process $\mathbf{Z}_t$, forms a VAR process. That is,

$$\begin{bmatrix}\mathbf{X}_t\\ \mathbf{Z}_t\end{bmatrix} = \begin{bmatrix}\mathbf{B} & \mathbf{C}\\ \mathbf{D} & \mathbf{E}\end{bmatrix} \cdot \begin{bmatrix}\mathbf{X}_{t-1}\\ \mathbf{Z}_{t-1}\end{bmatrix} + \boldsymbol{\varepsilon}_t, \tag{11}$$

where $\mathbf{Z}_t$ is not measured and can be considered as confounder time series, $\mathbf{B}$ is the causal transition matrix for the observed process $\mathbf{X}_t$, and $\mathbf{C}$ contains the the influence from $\mathbf{Z}_t$ to the observed process $\mathbf{X}_t$. The theoretical issue is whether $\mathbf{B}$ and $\mathbf{C}$ are identifiable from solely the observed process $\mathbf{X}_t$.

10.3.1 Practical Granger Causal Analysis Can Go Wrong

In practical Granger causal analysis, one just performs a linear regression of present on past on the observed $\mathbf{X}_t$ and then interpret the regression matrix causally. While making the ideal definition practically feasible, this may lead to wrong causal conclusions in the sense that it does not comply with the causal structure that we would infer given we had more information. Let us give an example for this. Let $\mathbf{X}_t$ be bivariate and $\mathbf{Z}_t$ be univariate. Moreover, assume

$$\begin{bmatrix}\mathbf{B} & \mathbf{C}\\ \mathbf{D} & \mathbf{E}\end{bmatrix} = \left(\begin{array}{cc|c} 0.9 & 0 & 0.5\\ 0.1 & 0.1 & 0.8\\ \hline 0 & 0 & 0.9\end{array}\right),$$

and let the covariance matrix of $\boldsymbol{\varepsilon}_t$ be the identity matrix. To perform practical Granger causal analysis, we proceed as usual: we fit a VAR model on *only* the observable process $\mathbf{X}_t$, in particular calculate the VAR transition matrix by

$$B_{pG} = \mathbb{E}(\mathbf{X}_t\mathbf{X}_{t-1}^\intercal)\mathbb{E}^{-1}(\mathbf{X}_t\mathbf{X}_t^\intercal) = \begin{pmatrix}0.89 & 0.35\\ 0.08 & 0.65\end{pmatrix}.$$

(up to rounding) and interpret the coefficients of B_{pG} as causal influences. Although, according to $\mathbf{B}$, the true time-delayed causal relations in $\mathbf{X}_t$, X_{2t} does not cause X_{1t}, B_{pG} suggests that there is a strong causal effect $X_{2,t-1} \to X_{1t}$ with the strength 0.35. It is even stronger than the relation $X_{1,t-1} \to X_{2t}$, which actually exists in the complete model with the strength 0.1.

10.3.2 Identifiability of $\mathbf{B}$ and Almost Identifiability of $\mathbf{C}$

Assume that all components of $\boldsymbol{\varepsilon}_t$ are non-Gaussian and that the dimensionality of the hidden process $\mathbf{Z}_t$ is not higher than that of the observed process $\mathbf{X}_t$. Together with some further technical assumptions, it has been shown that $\mathbf{B}$ is identifiable from $\mathbf{X}_t$; furthermore, the set of columns of $\mathbf{C}$ with at least two non-zero entries is identifiable from up to scaling of those columns [11].

One can then use a MoG to represent the distributions of the components of $\boldsymbol{\varepsilon}_t$ and develop a variation EM algorithm to estimate $\mathbf{B}$ and $\mathbf{C}$ from solely $\mathbf{X}_t$.

11. CONCLUSION AND OPEN PROBLEMS

We have reviewed central concepts in and fundamental methodologies for causal inference and discovery. The concepts include manipulations, causal models, sample predictive modeling, causal predictive modeling, structural equation models, the causal Markov assumption, and the faithfulness assumption. We have discussed the constraint-based causal structure search and its properties. In the second part of the paper, we have given a survey of structural equation models which enable us to fully identify causal structure from observational data. We focused on the two-variable case, where the task is to distinguish cause from effect. We have reviewed the linear non-Gaussian causal model, nonlinear additive noise model, and the post-nonlinear causal model, listed from the most to the least restrictive. We addressed the identifiablility of the causal direction: for those three models, in the generic case the backwards direction does not admit an independent error term and, as a consequence, it is possible to distinguish cause from effect. We have also briefly discussed the procedure to do so, which consists of fitting the structural equation model and doing independence test between the estimated error term and the hypothetical cause.

In the last three decades, enlightening progress has been made in the field of causal discovery and inference. However, there are still many fundamental questions to be answered:[8]

- What new models are appropriate for different combinations of kinds of data, e.g., experimental and observational [4, 7, 40, 8, 41, 9]?

- What new models are appropriate for different kinds of background knowledge, and different families of densities?

- What kind of scores can be used to best evaluate causal models from various kinds of data? In a related vein, what are good families of prior distributions that capture various kinds of background knowledge?

- How can search algorithms be improved to incorporate different kinds of background knowledge, search over different classes of causal models, run faster, handle more variables and larger sample sizes, be more reliable at small sample sizes, and produce output that is as informative as possible?

- For existing and novel causal search algorithms, what are their semantic and syntactic properties (e.g., soundness, consistency, maximum informativeness)? What are their statistical properties (pointwise consistency, uniform consistency, sample efficiency)? What are their computational properties (computational complexity)?

- What plausible alternatives are there to the Causal Markov and Faithfulness Assumptions? Are there other assumptions might be weaker and hold in more domains and applications without much loss about what can be reliably inferred? Are there stronger assumptions that are plausible for some domains that might allow for stronger causal inferences? How often are these assumption violated, and how much do violations of these assumptions lead to incorrect inferences?

- There are special assumptions, such as linearity, which can improve the strength of causal conclusions that can be reliably inferred, and the speed and sample efficiency of algorithms that draw the conclusions. What other distribution families or stronger assumptions about a domain are there that are plausible for some domains and how can they be used to improve causal inference?

- Can various statistical assumptions be relaxed? For example, what if the sample selection process is not i.i.d., but may be causally affected by variables of interest [2, 37, 5, 3, 30]?

In addition, there is also a number of open problems concerning SEM-based causal discovery and the asymmetry between cause and effect.

- First, one can consider structural equation models as a way to represent the conditional distribution of the effect given the cause. Can we then find hints as to the causal direction directly from the data distribution? In other words, can we find a general way to directly characterize the causal asymmetry in light of certain properties of the data distribution? If we managed to do so, it would hopefully put the causal Markov condition, the independent noise condition (in the SEMs), and the independent transformation condition in the nonlinear noiseless case [19] under the same umbrella. To this end, an attempt has been made by exploiting the so-called "exogeneity" property of a causally sufficient causal system [48]. But it is not clear whether this property is able to bring about computationally efficient and widely applicable causal discovery methods.

- Secondly, note that nonlinear structural equation models are usually intransitive. That is, if both causal processes $X_1 \to X_2$ and $X_2 \to X_3$ admit a particular type of structural equation model, say, the nonlinear additive noise model, the process $X_1 \to X_3$ does not necessarily follow the same model. (Linear models are transitive.) This could be a potential issue with structural-equation-model-based causal discovery: it may fail to discover indirect causal relations. (Here by direction causal relations, we mean the causal relations in which only a single noise variable is involved.) On the other hand, this may be a benefit of using structural equation models for causal discovery, in that it is possible to detect the existence of causal intermediate variables and further recover them. But how to do so is currently unclear.

- We have discussed how different types of independence, including conditional independence in the causal Markov condition and statistical independence between the error term and hypothetical cause in structural equations models, help to discovery causal information from data. On the other hand, it has been demonstrated that this type of independence (which is, loosely speaking, the independence between how the cause is generated and how the effect is generated from cause) is able to facilitate understanding and solving some machine learning or data analysis problems. For instance, it implies that when the feature causes the label (or target), unlabeled data points will *not* help in the semi-supervised learning scenario [31], and inspired new settings and formulations for domain adaptation by characterizing what information to transfer [43]. It is under investigation whether other machine learning

[8] The content and organization of the following open questions are largely due to suggestions from Constantin Aliferis, whom I thank for his suggestions.

methods including "adaptive boosting" can be understood from the causal perspective. In addition, it is unclear whether the learning guarantees for supervised learning actually depend on the causal relationship between the feature and target (or label), i.e., the causal role of the feature w.r.t. the target.

- Finally, developing efficient methods for causal discovery of more than two variables based on structural equation models is an important step towards large-scale causal analysis in various domains including neuroscience and biology. To make causal discovery computationally efficient, one may have to limit the complexity of the causal structure, say, limit the number of direct causes of each variable. Even so, a smart optimization procedure instead of exhaustive search is still missing in the literature.

12. REFERENCES

[1] P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Vol I.* Prentice Hall, 2000. 2nd Edition.

[2] G. F. Cooper. Causal discovery from data in the presence of selection bias. In *Fifth International Workshop on AI and Statistics*, pages 140–150, 1995.

[3] G. F. Cooper. A Bayesian method for causal modeling and discovery under selection. In *Uncertainty In Artificial Intelligence*, pages 98–106, 2000.

[4] G. F. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *Uncertainty In Artificial Intelligence*, pages 116–125, 1999.

[5] D. R. Cox and N. Wermuth. *Multivariate Dependencies: Models, Analysis and Interpretation (Monographs on Statistics and Applied Probability).* Chapman & Hall/CRC, 1996.

[6] H. Cramér. *Random variables and probability distributions.* Cambridge University Press, Cambridge, UK, 3rd edition, 1970.

[7] D. Danks. Learning the causal structure of overlapping variable sets. In *Lect Notes Comput Sc, 2534*, pages 178–191, 2002.

[8] F. Eberhardt, C. Glymour, and R. Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *21st Conference on Uncertainty in Artificial Intelligence*, pages 178–184, 2005.

[9] F. Eberhardt, C. Glymour, and R. Scheines. 4 n-1 experiments suffice to determine the causal relations among n variables. In D. E. Holmes and C. J. Lakhmi, editors, *Innovations in Machine Learning: Theory And Applications*, pages 97–112, 2006.

[10] F. Fisher. Ba correspondence principle for simultaneous equation models. *Econometrica*, 38:73–92, 1970.

[11] P. Geiger, K. Zhang, M. Gong, D. Janzing, and B. Schölkopf. Causal inference by identification of vector autoregressive processes with hidden components. In *Proc. 32th International Conference on Machine Learning (ICML 2015)*, 2015.

[12] M. Gong*, K. Zhang*, D. Tao, P. Geiger, and B. Schölkopf. Discovering temporal causal relations from subsampled data. In *Proc. 32th International Conference on Machine Learning (ICML 2015)*, 2015.

[13] C. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2, 1980.

[14] A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In S. Jain, H.-U. Simon, and E. Tomita, editors, *Algorithmic Learning Theory: 16th International Conference*, pages 63–78, Berlin, Germany, 2005. Springer.

[15] P. Hoyer, D. Janzing, J. Mooji, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, Vancouver, B.C., Canada, 2009.

[16] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley & Sons, Inc, 2001.

[17] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

[18] A. Hyvärinen, K. Zhang, S. Shimizu, and P. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, pages 1709–1731, 2010.

[19] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniuvsis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, pages 1–31, 2012.

[20] A. M. Kagan, Y. V. Linnik, and C. R. Rao. *Characterization Problems in Mathematical Statistics.* Wiley, New York, 1973.

[21] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H. G. Leimer. Independence properties of directed markov fields. *Networks*, 20:491–505, 1990.

[22] J. Lin. Factorizing multivariate function classes. In *Advances in Neural Information Processing Systems 10*, pages 563–569, Cambridge, MA, 1998. MIT Press.

[23] C. Meek. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 411–419, 1995.

[24] J. Mooij, O. Stegle, D. Janzing, K. Zhang, and B. Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, Curran, NY, USA, 2010.

[25] F. C. Palm and T. E. Nijman. Missing observations in the dynamic regression model. *Econometrica*, 52:1415–1435, 1984.

[26] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.

[27] J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge, 2000.

[28] J. Peters, J. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. In *Proc. UAI 2011*, pages 589–598, 2011.

[29] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning.* MIT Press, Cambridge,

Massachusetts, USA, 2006.

[30] T. Richardson and P. Spirtes. Ancestral graph markov models. *Ann Stat*, 30:962–1030, 2002.

[31] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proc. 29th International Conference on Machine Learning (ICML 2012)*, Edinburgh, Scotland, 2012.

[32] S. Shimizu, P. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

[33] C. A. Sims. Macroeconomics and reality. *Econometrica*, 48:1–48, 1980.

[34] P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:62–72, 1991.

[35] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* Spring-Verlag Lectures in Statistics, 1993.

[36] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT Press, Cambridge, MA, 2nd edition, 2001.

[37] P. Spirtes, C. Meek, and T. S. Richardson. Causal inference in the presence of latent variables and selection bias. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 499–505, 1995.

[38] P. Spirtes and R. Scheines. Causal inference of ambiguous manipulations. *Philos Sci,*, 71:833–845, 2004.

[39] A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Trans. on Signal Processing*, 47(10):2807–2820, 1999.

[40] C. Yoo and G. F. Cooper. An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. *Artif Intell Med*, 31:169–182, 2004.

[41] C. Yoo, G. F. Cooper, and M. Schmidt. A control study to evaluate a computer-based microarray experiment design recommendation system for gene-regulation pathways discovery. *J Biomed Inform*, 39:126–146, 2006.

[42] K. Zhang and L. Chan. Extensions of ICA for causality discovery in the hong kong stock market. In *Proc. 13th International Conference on Neural Information Processing (ICONIP 2006)*, 2006.

[43] K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 3150–3157. AAAI Press, 2015.

[44] K. Zhang and A. Hyvärinen. Causality discovery with additive disturbances: An information-theoretical perspective. In *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 2009*, Bled, Slovenia, 2009.

[45] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, 2009.

[46] K. Zhang, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, Barcelona, Spain, 2011.

[47] K. Zhang, Z. Wang, J. Zhang, and B. Schölkopf. On estimation of functional causal models: General results and application to post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technologies*, 2015. forthcoming.

[48] K. Zhang, J. Zhang, and B. Schölkopf. Distinguishing cause from effect based on exogeneity. In *Proc. 15th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2015)*, 2015.

Resolving Distributed Knowledge

Extended Abstract

Thomas Ågotnes
University of Bergen, Norway
thomas.agotnes@uib.no

Yì N. Wáng
Zhejiang University, China
ynw@zju.edu.cn

ABSTRACT

Distributed knowledge is the sum of the knowledge in a group; what someone who is able to discern between two possible worlds whenever *any* member of the group can discern between them, would know. Sometimes distributed knowledge is referred to as the potential knowledge of a group, or the joint knowledge they could obtain if they had unlimited means of communication. In epistemic logic, the formula $D_G\varphi$ is intended to express the fact that group G has distributed knowledge of φ, that there is enough information in the group to infer φ. But this is not the same as reasoning about *what happens if the members of the group share their information.* In this paper we introduce an operator R_G, such that $R_G\varphi$ means that φ is true after G have shared all their information with each other – after G's distributed knowledge has been *resolved.* The R_G operators are called *resolution operators*. Semantically, we say that an expression $R_G\varphi$ is true iff φ is true in what van Benthem [8, p. 249] calls (G's) *communication core*; the model update obtained by removing links to states for members of G that are not linked by *all* members of G. We study logics with different combinations of resolution operators and operators for common and distributed knowledge. Of particular interest is the relationship between distributed and common knowledge. The main results are sound and complete axiomatizations.

1. INTRODUCTION

In epistemic logic [3, 5, 12] different notions of *group knowledge* describe different ways in which knowledge can be associated with a group. *Common knowledge* is stronger than individual knowledge: that something is common knowledge requires not only that everybody in the group knows it, but that everybody knows that everybody knows it, and so on. *Distributed knowledge*, on the other hand, is weaker than individual knowledge: distributed knowledge is knowledge that is distributed throughout the group even if no individual knows it.

More concrete informal descriptions of the concept of distributed knowledge abound, but they are often inaccurate descriptions of the concept as formalized in standard epistemic logic. A misconception is that something is distributed knowledge in a group if the agents in the group could get to know it after some (perhaps unlimited) communications between them[1]. To see that this interpretation must be incorrect, consider the formula $D_{\{1,2\}}(p \wedge \neg K_1 p)$. In this formula, $D_G\varphi$ and $K_i\varphi$ mean that φ is distributed knowledge in the group G, and individual knowledge of agent i, respectively. Thus, the formula says that it is distributed knowledge among agents 1 and 2 that p is true and that agent 1 does not know p. This formula is *consistent* (also when we assume that knowledge has the S5 properties). However, it is not possible that agents 1 and 2 both can get to know that p is true and that agent 1 does not know that p is true (assuming the S5 properties of knowledge), no matter how much they communicate (or "pool" their knowledge). The "problem" here is that in a formula $D_G\psi$, ψ describes the possible states of the world as they were before any communication or other events took place, so a more accurate reading of $D_{\{1,2\}}(p\wedge\neg K_1 p)$ would perhaps be that it follows from the combination of 1 and 2's knowledge that $p \wedge \neg K_1 p$ *were true before any communication or other events took place*. More technically, the "problem" is due to the standard compositional semantics of modal logic: in the evaluation of $D_G\varphi$, the D_G operator picks out a number of states considered possible by the group G (actually the states considered possible by *all* members of the group), and then φ is evaluated in each of these states *in the original model, without any effect of the* D_G *operator*.

But we don't really consider this a problem. There are other interpretations of distributed knowledge where the consistency of the mentioned formula makes perfect sense, such that distributed knowledge being the knowledge of a third party, someone "outside the system" who somehow has access to the epistemic states of the group members. It shows, however, that it does not make sense to interpret distributed knowledge as something that is true after the agents in the group have communicated *with each other* – with the standard semantics.

In this paper we introduce and study an alternative group modality R_G, where $R_G\varphi$ means (roughly speaking) that φ is true after the agents in the group have shared all their information with each other. We call that *resolving* distributed knowledge, and the R_G operators are called *resolution operators*.

Semantically, we say that an expression $R_G\varphi$ is true iff φ is true in what van Benthem [8, p. 249] calls (G's) *communication core*; the model update obtained by removing links to states for members of G that are not linked by *all* members of G. See Fig. 1 for an illustration. In this paper we capture that model transformation by the new resolution operators, and study resulting logics. For example, the formula $R_{\{1,2\}}(p\wedge\neg K_1 p)$ will be inconsistent in the resulting logics. $R_{\{1,2\}}(p \wedge K_1 p)$ is true in state t in the model in Fig. 1.

This model transformation abstracts away from the issue of *how*

[1] Some examples of informal descriptions of distributed knowledge from the literature include "A group has distributed knowledge of a fact φ if the knowledge of φ is distributed among its members, so that by pooling their knowledge together the members of the group can deduce φ" [3]; ".. it should be possible for the members of the group to establish φ through communication" [11, 7]; ".. the knowledge that would result of the agents could somehow 'combine' their knowledge" [11]. These descriptions can at least give a reader the impression that distributed knowledge is about internal communication in the group of agents.

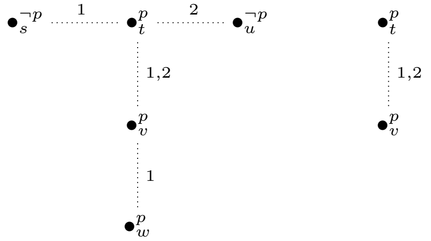


Figure 1: Example taken from [8, p. 248]. Model on the left, its communication core (for the set of all agents $\{1, 2\}$) on the right. Reflexivity, symmetry and transitivity are implicitly assumed.

the agents share their information; whether they *communicate* directly with each other and if so in which language, whether they are informed by some outsider about the information other agents have and if so how, and so on. As noted by van Benthem [8, p. 249], the communication core cannot always be obtained by *public announcements* using the epistemic language. Similarly, as noted by several researchers [11, 4, 7], standard distributed knowledge does not always follow logically from the knowledge of the individual agents expressible in the epistemic language. Our model, like that of standard distributed knowledge, is purely semantic: we assume that if an agent can discern between two different worlds, then there exists some mechanism that results in other members of the group being able to make the same distinction. This is further discussed Section 5.

This model transformation models a particular kind of internal group information sharing event. Exactly *which* kind depends on what we assume about what *other* agents, i.e., agents that are not in the group G that resolve their knowledge, know about the fact that this event is taking place. In this paper we will assume that it is common knowledge among the other agents that G resolve their knowledge – but not what the agents in G actually learn. This corresponds to a natural class of events: publicly observable private resolution of distributed knowledge. An example is a meeting in a closed room, where it is observed that a certain group meets in the room to share information.

We want to make it clear that we do not consider distributed knowledge with standard semantics to be “wrong”; the important thing is to be clear about its meaning. In particular, the resolution operators are not intended as a “replacement” of distributed knowledge operators, but as a complement: they express different things. The logics we study contains both types of operators, as well as common knowledge. The main results are sound and complete axiomatizations.

Technically, the model transformation, which amounts to removing certain edges, is similar to those found in the simplest dynamic epistemic logics [12] such as public announcement logics [6]. [8] has also pointed out the close connection between the communication core and sequences of public announcements. Public announcement logics with distributed knowledge have been studied recently [13]. In the absence of *common* knowledge, we get reduction axioms for public announcement logic with distributed knowledge. This turns out to be the case for resolution operators as well. It is not the case in the presence of common knowledge, however.

There is a close connection between the communication core and common knowledge [8]. By studying complete axiomatizations of logics with the resolution operators we make some aspects of that connection precise and give an answer to the question “when does distributed knowledge become common knowledge?” – under certain assumptions.

The rest of the paper is organized as follows. In the next section we review some background definitions and results from the literature, before we introduce logics with the new resolution operators in Section 3, where we also look at some properties of the operators. In Section 4 we prove completeness of resulting logics; the most interesting case being epistemic logic with common and distributed knowledge and resolution operators. We discuss related and future work and conclude in Section 5.

2. BACKGROUND

In this section we give a (necessarily brief) review of the main background concepts from the literature.

We henceforth assume a countable set of propositional variables $\textsc{prop}$ and a finite set of agents $\textsc{ag}$. We let $\textsc{gr}$ be the set of all non-empty groups, i.e., $\textsc{gr} = \wp(\textsc{ag}) \setminus \emptyset$.

An *epistemic model* over $\textsc{prop}$ and $\textsc{ag}$ (or just a *model*) $\mathfrak{M} = (S, \sim, V)$ where S is a set of states (or worlds), $V : \textsc{prop} \to 2^S$ associates a set of states $V(p)$ with each propositional variable p, and $\sim$ is a function that maps each agent to a binary equivalence relation on S. We write $\sim_i$ for $\sim(i)$.

$s \sim_i t$ means that agent i cannot discern between states s and t – if we are in s she doesn't know whether we are in t, and vice versa. Considering the distributed knowledge of a group G – a key concept in the following – we define a derived relation $\sim_G = \bigcap_{a\in G} \sim_a$ (it is easy to see that $\sim_G$ is an equivalence relation). Intuitively, someone who has all the knowledge of all the members of G can discern between two states if and only if at least one member of G can discern between them. We will also consider common knowledge. A similar relation modeling the common knowledge of a group is obtained by taking the transitive closure of the union of the individual relations: $\backsim_{C_G} = (\bigcup_{i\in G} \sim_i)^*$.

Definition 1. *Below are several languages from the literature.*

$$\begin{array}{ll}
(\mathcal{ELD}) & \varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid D_G\varphi \\
(\mathcal{ELCD}) & \varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid D_G\varphi \mid C_G\varphi \\
(\mathcal{PACD}) & \varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid D_G\varphi \mid C_G\varphi \mid [\varphi]\varphi,
\end{array}$$

where $p \in \textsc{prop}$, $i \in \textsc{ag}$ and $G \in \textsc{gr}$. We use the usual propositional derived operators, as well as $E_G\varphi$ for $\bigwedge_{i\in G} K_i\varphi$.

$\mathcal{ELD}$ and $\mathcal{ELCD}$ are static epistemic languages with distributed knowledge, and with distributed and common knowledge, respectively. These are the languages we will extend with resolution operators in the next section. We will also be interested in $\mathcal{PACD}$, the language for public announcement logic with both common knowledge and distributed knowledge, when we look at completeness proofs.

Satisfaction of a formula φ of any of these languages in a state m of a model $\mathfrak{M}$, denoted $\mathfrak{M}, m \models \varphi$, is defined recursively by the following clauses:

$$\begin{array}{lll}
\mathfrak{M}, m \models p & \text{iff} & m \in V(p) \\
\mathfrak{M}, m \models \neg\varphi & \text{iff} & \mathfrak{M}, m \not\models \varphi \\
\mathfrak{M}, m \models \varphi \wedge \psi & \text{iff} & \mathfrak{M}, m \models \varphi \;\&\; \mathfrak{M}, m \models \psi \\
\mathfrak{M}, m \models K_a\varphi & \text{iff} & \forall n \in S.\ (m \sim_a n \Rightarrow \mathfrak{M}, n \models \varphi) \\
\mathfrak{M}, m \models D_G\varphi & \text{iff} & \forall n \in S.\ (m \sim_G n \Rightarrow \mathfrak{M}, n \models \varphi) \\
\mathfrak{M}, m \models C_G\varphi & \text{iff} & \forall n \in S.\ (m(\bigcup_{i\in G} \sim_i)^* n \Rightarrow \mathfrak{M}, n \models \varphi) \\
\mathfrak{M}, m \models [\psi]\varphi & \text{iff} & \mathfrak{M}, m \models \psi \Rightarrow \mathfrak{M}|\psi, m \models \varphi.
\end{array}$$

where R^* denotes the transitive closure of R and $\mathfrak{M}|\psi$ is the submodel of $\mathfrak{M}$ restricted to $\{m \in M \mid \mathfrak{M}, m \models \psi\}$. *Validity* is defined as usual: $\models \varphi$ means that $\mathfrak{M}, m \models \varphi$ for all $\mathfrak{M}$ and m.

We now define some axiom schemata and rules. The classical "S5" proof system for multi-agent epistemic logic, denoted (**S5**), consists of the following axioms and rules:

(PC)	instances of tautologies
(K)	$K_i(\varphi \to \psi) \to (K_i\varphi \to K_i\psi)$
(T)	$K_i\varphi \to \varphi$
(4)	$K_i\varphi \to K_iK_i\varphi$
(5)	$\neg K_i\varphi \to K_i\neg K_i\varphi$
(MP)	from φ and $\varphi \to \psi$ infer ψ
(N)	from φ infer $K_i\varphi$.

Axioms for distributed knowledge, denoted (**DK**):

(K$_D$)	$D_G(\varphi \to \psi) \to (D_G\varphi \to D_G\psi)$
(T$_D$)	$D_G\varphi \to \varphi$
(5$_D$)	$\neg D_G\varphi \to D_G\neg D_G\varphi$
(D1)	$K_i\varphi \leftrightarrow D_i\varphi$
(D2)	$D_G\varphi \to D_H\varphi$, if $G \subseteq H$.

Axioms and rules for common knowledge, denoted (**CK**):

(K$_C$)	$C_G(\varphi \to \psi) \to (C_G\varphi \to C_G\psi)$
(T$_C$)	$C_G\varphi \to \varphi$
(C1)	$C_G\varphi \to E_GC_G\varphi$
(C2)	$C_G(\varphi \to E_G\varphi) \to (\varphi \to C_G\varphi)$
(N$_C$)	from φ infer $C_G\varphi$.

The system that consists of (**S5**) and (**DK**) over the language $\mathcal{ELD}$, denoted **S5D**, is a sound and complete axiomatization of all $\mathcal{ELD}$ validities. The system that consists of (**S5**), (**DK**) and (**CK**) over the language $\mathcal{ELCD}$ is a sound and complete axiomatization of all $\mathcal{ELCD}$ validities.

3. RESOLVING DISTRIBUTED KNOWLEDGE

We want to model the event that G resolves their knowledge. An immediate question is: whenever the group G is a proper subset of the set of all agents, what do *the other* agents know about the fact that this event takes place? Here we will model situations where it is common knowledge among the other agents *that* the event takes place, but not *what* the members of the group learn. As discussed in the introduction, this corresponds to a natural class of information sharing events, namely publicly observable private communication, such as a meeting in a closed room that is observed to be taking place. This is captured by a *global* model update: in every state, remove a link to another state for any member of G whenever it is not the case that there is a link to that state for *all* members of G.

Formally, given a model $\mathfrak{M} = (S, \sim, V)$ and a group of agents G, the *(global) G-resolved update of* $\mathfrak{M}$ is the model $\mathfrak{M}|_G$ where $\mathfrak{M}|_G = (S, \sim|_G, V)$ and

$$(\sim|_G)_i = \begin{cases} \bigcap_{j \in G} \sim_j, & i \in G, \\ \sim_i, & \text{otherwise.} \end{cases}$$

We consider the following new languages with resolution operators.

Definition 2 (Languages).

$$\begin{array}{ll} (\mathcal{RD}) & \varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid D_G\varphi \mid R_G\varphi \\ (\mathcal{RCD}) & \varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid D_G\varphi \mid C_G\varphi \mid R_G\varphi, \end{array}$$

where $p \in \text{PROP}$, $i \in \text{AG}$ *and* $G \in \text{GR}$.

The interpretation of these languages in a pointed model is defined as usual, with the following additional clause for the resolution operator:

$$\mathfrak{M}, s \models R_G\varphi \quad \text{iff} \quad \mathfrak{M}|_G, s \models \varphi.$$

A couple of observations. Recall that we write $\sim_H$ for $\bigcap_{i \in H} \sim_i$. Thus,

$$(\sim|_G)_i = \begin{cases} \sim_G, & i \in G, \\ \sim_i, & i \notin G. \end{cases} \quad (\sim|_G)_H = \begin{cases} \sim_H, & G \cap H = \emptyset, \\ \sim_{G \cup H}, & G \cap H \neq \emptyset. \end{cases}$$

Also note that $(\sim|_G)_i = (\sim|_G)_{\{i\}}$.

3.1 Some Validities

Let us start with a trivial validity: resolution has no effect for a singleton coalition.

Proposition 3. *The following is valid, where* $i \in \text{AG}$ *and* $\varphi \in \mathcal{RCD}$: $R_{\{i\}}\varphi \leftrightarrow \varphi$.

More interesting are the following properties.

Proposition 4 (Reduction Principles). *The following are valid, where* $G, H \in \text{GR}$, $p \in \text{PROP}$ *and* $\varphi \in \mathcal{RCD}$:

1. $R_Gp \leftrightarrow p$
2. $R_G(\varphi \wedge \psi) \leftrightarrow R_G\varphi \wedge R_G\psi$
3. $R_G\neg\varphi \leftrightarrow \neg R_G\varphi$
4. $R_GK_i\varphi \leftrightarrow D_GR_G\varphi$, *when* $i \in G$
5. $R_GK_i\varphi \leftrightarrow K_iR_G\varphi$, *when* $i \notin G$
6. $R_GD_H\varphi \leftrightarrow D_{G\cup H}R_G\varphi$, *when* $G \cap H \neq \emptyset$
7. $R_GD_H\varphi \leftrightarrow D_HR_G\varphi$, *when* $G \cap H = \emptyset$.

These properties are reduction principles, of the type known from public announcement logic: they allow us to simplify expressions involving resolution operators. If we have such principles for the combination of resolution with all other operators we can eliminate resolution operators altogether. There are two cases missing above: R_GC_H and R_GR_H[2]. We consider them next.

3.1.1 Common Knowledge

First, after the *grand* coalition have resolved their knowledge, then all the distributed information in the system is common knowledge: there is no longer a distinction between distributed and common knowledge:

Proposition 5. *For any* $\varphi \in \mathcal{RCD}$: $R_{\mathrm{AG}}C_{\mathrm{AG}}\varphi \leftrightarrow R_{\mathrm{AG}}D_{\mathrm{AG}}\varphi$.

Proof. 5. Given a model $\mathfrak{M} = (S, \sim, V)$ and $s \in S$,

$$\begin{array}{ll} & \mathfrak{M}, s \models R_{\mathrm{AG}}C_{\mathrm{AG}}\varphi \\ \text{iff} & \mathfrak{M}|_{\mathrm{AG}}, s \models C_{\mathrm{AG}}\varphi \\ \text{iff} & \forall t \in S.\ (s(\sim|_{\mathrm{AG}})_{C_{\mathrm{AG}}}t \Rightarrow \mathfrak{M}|_{\mathrm{AG}}, s \models \varphi) \\ \text{iff} & \forall t \in S.\ (s \sim_{\mathrm{AG}} t \Rightarrow \mathfrak{M}|_{\mathrm{AG}}, s \models \varphi) \quad (\dagger) \\ \text{iff} & \mathfrak{M}|_{\mathrm{AG}}, s \models D_{\mathrm{AG}}\varphi \\ \text{iff} & \mathfrak{M}, s \models R_{\mathrm{AG}}D_{\mathrm{AG}}\varphi, \end{array}$$

where for (†) we show that $(\sim|_{\mathrm{AG}})_{C_{\mathrm{AG}}} =\sim_{\mathrm{AG}}$. This is easy: by definition we can verify that for all $i \in \mathrm{AG}$, $(\sim|_{\mathrm{AG}})_i =\sim_{\mathrm{AG}}$; hence $(\sim|_{\mathrm{AG}})_{C_{\mathrm{AG}}} = (\bigcup_{i \in \mathrm{AG}}(\sim|_{\mathrm{AG}})_i)^* = (\sim_{\mathrm{AG}})^* =\sim_{\mathrm{AG}}$. □

For the general case, as in the case of distributed knowledge, we have that the resolution operators and common knowledge operators commute when the groups are *disjoint*:

Proposition 6. *Let i be an agent, G and H groups of agents and* $\varphi \in \mathcal{RCD}$. *The following hold:*

1. *If* $G \cap H = \emptyset$, *then* $\models R_GC_H\varphi \leftrightarrow C_HR_G\varphi$

[2] The lack of a reduction axiom for the general $R_GR_H\varphi$ case does not mean we cannot get a reduction in the language $\mathcal{RD}$: we can simply do the reduction "inside-out".

2. *If $G \supseteq H$ and $i \in G$, then* $\models R_G C_H \varphi \leftrightarrow R_G K_i \varphi \leftrightarrow D_G R_G \varphi$.

PROOF. See the appendix. □

However, this does not hold for overlapping groups G and H. In general, we have that (see the proof of the proposition above) $\mathfrak{M}, s \models R_G C_H \varphi$ iff $\mathfrak{M}|_G, t \models \varphi$ for any $(s,t) \in \sim_H^{*'}$, where $\sim_H^{*'} = (\bigcap_{i \in G} \sim_i \cup \bigcup_{i \in H \setminus G} \sim_i)^*$. This does not seem to be reducible.

3.1.2 Iterated resolution

What about $R_G R_H \varphi$? In extreme cases, we have:

PROPOSITION 7. *The following are valid, where $G, H \in$ GR and $\varphi \in \mathcal{RCD}$:*

1. $R_G R_H \varphi \leftrightarrow R_H R_G \varphi$, *if $G \cap H = \emptyset$*
2. $R_G R_G \varphi \leftrightarrow R_G \varphi$.

However, in the general case there does not seem to be a reduction axiom in this case. In particular, $R_G R_H \varphi$ is not equivalent to $R_{G \cup H} \varphi$.

Let us consider an example of iterated resolution.

EXAMPLE 8 (TRIPLE UPDATE). *Let $\mathfrak{M} = (S, \sim, V)$ and $\mathfrak{M}|_{G_1}|_{G_2}|_{G_3} = (S, \sim|_{G_1}|_{G_2}|_{G_3}, V)$. For any agent i, for any number x, we write G_x for "$i \in G_x$", and $\overline{G_x}$ for "$i \notin G_x$". Then*

$$(\sim|_{G_1}|_{G_2}|_{G_3})_i = \begin{cases} \text{if } \overline{G_1 G_2 G_3}: & \sim_i \\ \text{if } G_1 \overline{G_2 G_3}: & \sim_{G_1} \\ \text{if } \overline{G_1} G_2 \overline{G_3} \begin{cases} G_1 \cap G_2 = \emptyset: \\ G_1 \cap G_2 \neq \emptyset: \end{cases} & \begin{matrix} \sim_{G_2} \\ \sim_{G_1 \cup G_2} \end{matrix} \\ \text{if } G_1 G_2 \overline{G_3}: & \sim_{G_1 \cup G_2} \\ \text{if } \overline{G_1 G_2} G_3: & \sim_{G_3} \\ \text{if } G_1 \overline{G_2} G_3: & \sim_{G_1 \cup G_3} \\ \text{if } \overline{G_1} G_2 G_3 \begin{cases} G_1 \cap G_2 = \emptyset: \\ G_1 \cap G_2 \neq \emptyset: \end{cases} & \begin{matrix} \sim_{G_2 \cup G_3} \\ \sim_{G_1 \cup G_2 \cup G_3} \end{matrix} \\ \text{if } G_1 G_2 G_3: & \sim_{G_1 \cup G_2 \cup G_3} \end{cases}$$

In general we get the following (the proof is straightforward from the semantic definition).

PROPOSITION 9. *Let $M = (S, \sim, V)$ and $M|_{G_1}|\cdots|_{G_n} = (S, \sim|_{G_1}|\cdots|_{G_n}, V)$. Then, following the notation of Example 8, for any $i \in$ AG,*

$$(\sim|_{G_1}|\cdots|_{G_n})_i = \begin{cases} \sim_i, & \text{if } \overline{G_1 \cdots G_n} \\ \sim_{G_1 \cup \Theta}, & \text{if starting with } \overline{G_1} G_2 \text{ and} \\ & G_1 \cap G_2 \neq \emptyset \\ \sim_\Theta, & \text{otherwise} \end{cases}$$

where Θ is the union of all G_x such that $i \in G_x$.

3.2 Reduction Normal Form for Individual and Distributed Knowledge

As we see from the previous section, the reduction axioms for individual knowledge and distributed knowledge both contain two distinct cases, and the principles of iterated resolution become more complicated. In this section we give a unique form for such reductions, which will be of use later when we prove completeness. We shall call it *reduction normal form* for individual and distributed knowledge.

DEFINITION 10 (δ FUNCTION). *Given an agent i, a group H, and a sequence of groups $G_1, \ldots, G_n$, we define a function δ as follows:*

$$\delta_0 = \begin{cases} G_n \cup H, & G_n \cap H \neq \emptyset \\ H, & G_n \cap H = \emptyset \end{cases}$$

$$\delta_x = \begin{cases} G_{n-x} \cup \delta_{x-1}, & G_{n-x} \cap \delta_{x-1} \neq \emptyset \\ \delta_{x-1}, & G_{n-x} \cap \delta_{x-1} = \emptyset \end{cases}$$

$$\delta(H, G_1, \ldots, G_n) = \quad \delta_n.$$

Clearly $\delta(H, G_1, \ldots, G_n) \subseteq H \cup G_1 \cup \cdots \cup G_n$. We simply write δ instead of $\delta(H, G_1, \ldots, G_n)$ when its parameters are clear in the context.

PROPOSITION 11. *Let $i \in$ AG, $G_1, \ldots, G_n, H \in$ GR, $M = (S, \sim, V)$ and $M|_{G_1}|\cdots|_{G_n} = (S, \sim|_{G_1}|\cdots|_{G_n}, V)$. Then,*

1. $\models R_{G_1} \cdots R_{G_n} K_i \varphi \leftrightarrow D_{\delta(\{i\}, G_1, \ldots, G_n)} R_{G_1} \cdots R_{G_n} \varphi$;
2. $\models R_{G_1} \cdots R_{G_n} D_H \varphi \leftrightarrow D_{\delta(H, G_1, \ldots, G_n)} R_{G_1} \cdots R_{G_n} \varphi$;
3. $(\sim|_{G_1}|\cdots|_{G_n})_i = \sim_{\delta(\{i\}, G_1, \ldots, G_n)}$;
4. $(\sim|_{G_1}|\cdots|_{G_n})_H = \sim_{\delta(H, G_1, \ldots, G_n)}$.

PROOF. Straightforward: the recursive steps in the definition of the δ function matches exactly the reduction axioms. Note that clauses 1 and 3 can be treated as special cases of clauses 2 and 4 respectively. □

4. AXIOMATIZATIONS

We construct sound and complete axiomatizations of the logics for the two languages $\mathcal{RD}$ and $\mathcal{RCD}$.

4.1 Resolution and Distributed Knowledge

Consider the language $\mathcal{RD}$. Let **RD** be the system defined in Figure 2, where (**S5**) and (**DK**) are found in Section 2 and (**RR**) stands for the following reduction axioms for resolution:

(RA)	$R_G p \leftrightarrow p$
(RC)	$R_G(\varphi \wedge \psi) \leftrightarrow R_G \varphi \wedge R_G \psi$
(RN)	$R_G \neg \varphi \leftrightarrow \neg R_G \varphi$
(RD1)	$R_G D_H \varphi \leftrightarrow D_{G \cup H} R_G \varphi$, if $G \cap H \neq \emptyset$
(RD2)	$R_G D_H \varphi \leftrightarrow D_H R_G \varphi$, if $G \cap H = \emptyset$.

Note that (**RR**) contains most of the validities in Proposition 4, except for the reduction principles for individual knowledge – they are provable with RD1, RD2 and D1. In addition, we need the rule N_R for making a reduction to **S5D**. With the rule N_R we can easily show that the rule of *Replacement of Equivalents (RoE)* is admissible in **RD**. RoE allows us to carry out a reduction even without having a reduction axiom for iterated resolution.

(**S5**)	classical proof system for multi-agent epistemic logic
(**DK**)	characterization axioms for distributed knowledge
(**RR**)	reduction axioms for resolution
(N_R)	from φ infer $R_G \varphi$

Figure 2: Axiomatization RD.

THEOREM 12. *Any $\mathcal{RD}$ formula is valid if and only if it is provable in **RD**.*

4.2 Resolution, Distributed and Common Knowledge

Consider the language $\mathcal{RCD}$. Let **RCD** be the system defined in Figure 3, which extends **RD** with (**CK**), found in Section 2, and an induction rule for resolved common knowledge (RR_C).

(**S5**)	classical proof system for multi-agent epistemic logic
(**CK**)	axioms and rules for common knowledge
(**DK**)	characterization axioms for distributed knowledge
(N$_R$)	from φ infer $R_G\varphi$
(**RR**)	reduction axioms for resolution
(RR$_C$)	from $\varphi \to (E_H\varphi \wedge R_{G_1} \cdots R_{G_n}\psi)$ infer $\varphi \to R_{G_1} \cdots R_{G_n} C_H\psi$

Figure 3: Axiomatization RCD

4.2.1 Soundness

For soundness it suffices to show that the rule RR$_C$ preserves validity (we know that the other axioms/rules are valid/validity preserving from soundness results for the logics based on the sublanguages of $\mathcal{RCD}$).

LEMMA 13 (RR$_C$-VALIDITY PRESERVATION). *For all $\mathcal{RCD}$ formulas φ and ψ, all $G_1, \ldots, G_n, H \in$ GR, if $\models \varphi \to (E_H\varphi \wedge R_{G_1} \cdots R_{G_n}\psi)$, then $\models \varphi \to R_{G_1} \cdots R_{G_n} C_H\psi$.*

PROOF. Suppose $\models \varphi \to (E_H\varphi \wedge R_{G_1} \cdots R_{G_n}\psi)$. Given a model $\mathfrak{M}$ and a state s, suppose $\mathfrak{M}, s \models \varphi$, we must show that $\mathfrak{M}, s \models R_{G_1} \cdots R_{G_n} C_H\psi$, i.e., $\mathfrak{M}|_{G_1}| \cdots |_{G_n}, s \models C_H\psi$. Thus, for all H-paths $s_0(\sim |_{G_1}| \cdots |_{G_n})_{i_0} \cdots (\sim |_{G_1}| \cdots |_{G_n})_{i_{x-1}} s_x$, where $s = s_0$, we need to show that $\mathfrak{M}|_{G_1}| \cdots |_{G_n}, s_x \models \psi$.

From $\models \varphi \to (E_H\varphi \wedge R_{G_1} \cdots R_{G_n}\psi)$ and $\mathfrak{M}, s_0 \models \varphi$ we get $\mathfrak{M}, s_0 \models (E_H\varphi \wedge R_{G_1} \cdots R_{G_n}\psi)$, which entails:

$$\mathfrak{M}, s_1 \models \varphi \quad \text{and} \quad \mathfrak{M}|_{G_1}| \cdots |_{G_n}, s_0 \models \psi.$$

From $\mathfrak{M}, s_1 \models \varphi$ we get $\mathfrak{M}, s_1 \models (E_H\varphi \wedge R_{G_1} \cdots R_{G_n}\psi)$, which entails:

$$\mathfrak{M}, s_2 \models \varphi \quad \text{and} \quad \mathfrak{M}|_{G_1}| \cdots |_{G_n}, s_1 \models \psi.$$

By similar reasoning, for all $y = 0, \ldots, x$, we have

$$\mathfrak{M}, s_y \models \varphi \quad \text{and} \quad \mathfrak{M}|_{G_1}| \cdots |_{G_n}, s_y \models \psi,$$

which entails $\mathfrak{M}|_{G_1}| \cdots |_{G_n}, s_x \models \psi$ as we wish to show. □

COROLLARY 14 (SOUNDNESS). *For any $\mathcal{RCD}$ formula φ, if φ is provable in **RCD**, then it is valid.*

4.2.2 Completeness

As already discussed, **RCD** is similar to **PACD** (axiomatization for public announcement logic with common and distributed knowledge; see [13]): both logics extend epistemic logic with common and distributed knowledge with dynamic operators with update semantics that remove states. There does not seem, however, to be a trivial relationship between the two types of dynamic operators. We are nevertheless able to make heavy use of the completeness proof of **PACD** in [13] when proving completeness of **RCD**. That proof is again based on the completeness proof for public announcement logic with (only) common knowledge found in [2, 12], extended to deal with the distributed knowledge operators (which is non-trivial since intersection is not modally definable). In the following completeness proof we tweak the **PACD** proof to deal with resolution operators instead of public announcement operators. The general proof strategy is as follows: define a finite canonical pseudo model, where distributed knowledge operators are taken as primitive, and then transform it to a proper model while preserving truth. For the last step we can use a transformation based on *unraveling and folding* in [13] directly.

The most important difference to the **PACD** completeness proof in [13], and indeed the crux of the proof, is the use of the induction rule for resolved common knowledge (RR$_C$). No corresponding rule is needed in the **PACD** completeness proof. The rule is used in the proof of Lemma 29(8).

Pseudo Semantics.

DEFINITION 15 (PRE-MODELS[13]). *A* pre-model *is a tuple $\mathfrak{M} = (S, \backsim, V)$ where:*

- *S is a non-empty set of states;*
- *$\backsim$ is a function which maps every agent and every non-empty group of agents to an equivalence relation; we write $\backsim_i$ and $\backsim_G$ for $\backsim(i)$ and $\backsim(G)$ respectively;*
- *V : PROP $\to \wp(S)$ is a valuation.*

$\backsim_{C_G}$ is defined as the reflexive transitive closure of $\bigcup_{i \in G} \backsim_i$, just as for a model.

A pre-model is technically a model with a bigger set of agents (all groups are treated as agents in a pre-model). More precisely, if we make the set of agents A explicit in a pre-model, e.g., $\mathfrak{M} = (A, S, \backsim, V)$, then $\mathfrak{M}$ is in fact a "genuine" model $(S, \backsim, V)$ where the set of agents is $A \cup (\wp(A) \setminus \emptyset)$.

DEFINITION 16 (PSEUDO MODELS[13]). *A* pseudo model *is a pre-model $\mathfrak{M} = (S, \backsim, V)$ such that for any agent i and any groups G and H,*

- *$\backsim_{\{i\}} = \backsim_i$, and*
- *$G \subseteq H$ implies $\backsim_H \subseteq \backsim_G$.*

A pointed pre-model *(resp.* pointed pseudo model*) is a tuple $(\mathfrak{M}, s)$ consisted of a pre-model (resp. pseudo model) $\mathfrak{M}$ and a state s in $\mathfrak{M}$.*

DEFINITION 17 (PSEUDO SEMANTICS). *Given a pre-model $\mathfrak{M} = (S, \backsim, V)$, let m be a state in M.* Satisfaction at $(\mathfrak{M}, s)$ *is defined as follows:*

$$\begin{array}{lll}
\mathfrak{M}, s \models_{\mathsf{p}} p & \textit{iff} & s \in V(p) \\
\mathfrak{M}, s \models_{\mathsf{p}} \neg\varphi & \textit{iff} & \mathfrak{M}, s \not\models_{\mathsf{p}} \varphi \\
\mathfrak{M}, s \models_{\mathsf{p}} \varphi \wedge \psi & \textit{iff} & \mathfrak{M}, s \models_{\mathsf{p}} \varphi \;\&\; \mathfrak{M}, s \models_{\mathsf{p}} \psi \\
\mathfrak{M}, s \models_{\mathsf{p}} K_i\varphi & \textit{iff} & (\forall t \in \mathfrak{M})(s \backsim_i n \Rightarrow \mathfrak{M}, t \models_{\mathsf{p}} \varphi) \\
\mathfrak{M}, s \models_{\mathsf{p}} C_G\varphi & \textit{iff} & (\forall t \in \mathfrak{M})(s \backsim_{C_G} t \Rightarrow \mathfrak{M}, t \models_{\mathsf{p}} \varphi) \\
\mathfrak{M}, s \models_{\mathsf{p}} D_G\varphi & \textit{iff} & (\forall t \in \mathfrak{M})(s \backsim_G t \Rightarrow \mathfrak{M}, t \models_{\mathsf{p}} \varphi) \\
\mathfrak{M}, s \models_{\mathsf{p}} R_G\psi & \textit{iff} & \mathfrak{M}|_G, s \models_{\mathsf{p}} \varphi,
\end{array}$$

where $\mathfrak{M}|_G = (S, \backsim|_G, V)$ such that

$$(\backsim|_G)_i = \begin{cases} \backsim_G, i \in G \\ \backsim_i, \; i \notin G \end{cases} \text{ and } (\backsim|_G)_H = \begin{cases} \backsim_{H \cup G}, & H \cap G \neq \emptyset \\ \backsim_H, & H \cap G = \emptyset \end{cases}$$

Satisfaction in a pre-model $\mathfrak{M}$ *(denoted by $\mathfrak{M} \models_{\mathsf{p}} \varphi$) is defined as usual. We use $\models_{\mathsf{p}} \varphi$ to denote validity, i.e. $\mathfrak{M}, s \models_{\mathsf{p}} \varphi$ for any pointed pre-model $(\mathfrak{M}, s)$. We write $\models$ instead of $\models_{\mathsf{p}}$ when there is no confusion.*

PROPOSITION 18. *Let $\mathfrak{M}$ be a pseudo model, G a group of agents. Then $\mathfrak{M}|_G$ is a pseudo model.*

PROOF. See the appendix. □

PROPOSITION 19. *Propositions 9 and 11 still hold for pseudo models.*

When we regard a pre-model as a genuine model, classical (individual) bisimulation becomes an invariance relation. To make this clear, we first elaborate the definition of bisimulation for pre-models, and then introduce its invariance results.

Definition 20 (Pre-model bisimulation). *Let two pre-models $\mathfrak{M} = (S, \backsim, V)$ and $\mathfrak{M}' = (S', \backsim', V')$ be given. A non-empty relation $Z \subseteq S \times S'$ is called a* bisimulation *between $\mathfrak{M}$ and $\mathfrak{M}'$, denoted by $\mathfrak{M} \rightleftarrows \mathfrak{M}'$, if for all $\tau \in \text{AG} \cup \text{GR}$, all $s \in S$ and $s' \in S'$ such that sZs', the following hold.*

(at) *For all $p \in \text{PROP}$, $s \in V(p)$ iff $s' \in V'(p)$;*

(zig) *For all $t \in S$, if $s \sim_\tau t$, there is a $t' \in S'$ such that $s' \sim'_\tau t'$ and tZt';*

(zag) *For all $t' \in S'$, if $s' \sim'_\tau t'$, there is a $t \in S$ such that $s \sim_\tau t$ and tZt'.*

We say that pointed pre-models $(\mathfrak{M}, s)$ and $(\mathfrak{M}', s')$ are bisimilar, *denoted $(\mathfrak{M}, s) \rightleftarrows (\mathfrak{M}', s')$, if there is a bisimulation Z between $\mathfrak{M}$ and $\mathfrak{M}'$ linking s and s'.*

Proposition 21.*Resolution preserves pre-model bisimulation. I.e., for all pointed pre-models $(\mathfrak{M}, s)$ and $(\mathfrak{M}', s')$, if $(\mathfrak{M}, s) \rightleftarrows (\mathfrak{M}', s')$ then $(\mathfrak{M}|_G, s) \rightleftarrows (\mathfrak{M}'|_G, s')$.*

Proof. See the appendix. □

Corollary 22. *For any pre-models $\mathfrak{M}$ and $\mathfrak{M}'$, if $(\mathfrak{M}, s) \rightleftarrows (\mathfrak{M}', s')$ then $\mathfrak{M}, s \models_{\mathsf{P}} \varphi$ iff $\mathfrak{M}', s' \models_{\mathsf{P}} \varphi$ for any $\mathcal{RCD}$ formula φ.*

As introduced in [13], we can also consider a kind of bisimulations between genuine models and pre-models.

Definition 23 (Trans-bisimulation [13]). *Let a model $\mathfrak{M} = (M, \sim, V)$ and a pre-model $\mathfrak{N} = (N, \backsim, \nu)$ be given. A non-empty binary relation $Z \subseteq M \times N$ is called a* trans-bisimulation between $\mathfrak{M}$ and $\mathfrak{N}$, *if for all $m \in M$ and $n \in N$ with mZn:*

(at) *$m \in V(p)$ iff $n \in \nu(p)$ for all $p \in \text{PROP}$,*

(zig$_{ag}$) *For all $m' \in M$ and all $i \in \text{AG}$, if $m \sim_i m'$ (and so $m \sim_{\{i\}} m'$), then there is an $n' \in N$ such that $m'Zn'$ and $n \backsim_{\tau_0} \cdots \backsim_{\tau_x} n'$ with each of $\tau_0, \ldots, \tau_x$ being "i" or "G" such that $i \in G$;*

(zig$_{gr}$) *For all $m' \in M$ and all $G \in \text{GR}$ with $|G| \geq 2$, if $m \sim_G m'$, then there is an $n' \in N$ such that $m'Zn'$ and $n \backsim_{G_1} \cdots \backsim_{G_x} n'$ with $G \subseteq G_1 \cap \cdots \cap G_x$;*

(zag) *For all $n' \in N$ and all $\tau \in \text{AG} \cup \text{GR}$, if $n \backsim_\tau n'$, then there is an $m' \in M$ such that $m'Zn'$ and $m \sim_\tau m'$.*

We write $Z : (\mathfrak{M}, m) \rightleftarrows^{\text{T}} (\mathfrak{N}, n)$ if Z is a trans-bisimulation between $\mathfrak{M}$ and $\mathfrak{N}$ linking m and n. We say a pointed model $(\mathfrak{M}, m)$ and a pointed pre-model $(\mathfrak{N}, n)$ are trans-bisimilar, *denoted by $(\mathfrak{M}, m) \rightleftarrows^{\text{T}} (\mathfrak{N}, n)$, if there is a trans-bisimulation Z such that $Z : (\mathfrak{M}, m) \rightleftarrows^{\text{T}} (\mathfrak{N}, n)$.*

To make the notation symmetric, we call Z a trans-bisimulation between $\mathfrak{N}$ and $\mathfrak{M}$ *if it is a trans-bisimulation between $\mathfrak{M}$ and $\mathfrak{N}$, and we regard $Z : (\mathfrak{N}, n) \rightleftarrows^{\text{T}} (\mathfrak{M}, m)$ just as $Z : (\mathfrak{M}, m) \rightleftarrows^{\text{T}} (\mathfrak{N}, n)$.*

(Pseudo) satisfaction of $\mathcal{RCD}$ formulas is invariant under trans-bisimulation. We will not prove that directly at this point: it follows from a stronger result we prove later (Lemma 33).

Theorem 24 (Pseudo soundness). *All theorems of* ***RCD*** *are valid in the class of all pseudo models.*

Proof. See the appendix. □

Finitary Canonical Models.

Definition 25 (Closure). *Given a formula φ, the* closure *of φ is given by the function $cl : \mathcal{RCD} \to \wp(\mathcal{RCD})$ which is defined as follows:*

1. *$\varphi \in cl(\varphi)$, and if $\psi \in cl(\varphi)$, so are all of its subformulas;*
2. *If φ is not a negation, then $\varphi \in cl(\varphi)$ implies $\neg\varphi \in cl(\varphi)$;*
3. *$K_i\psi \in cl(\varphi)$ iff $D_{\{i\}}\psi \in cl(\varphi)$;*
4. *$C_G\psi \in cl(\varphi)$ implies $\{K_iC_G\psi \mid a \in A\} \subseteq cl(\varphi)$;*
5. *$R_{G_1} \cdots R_{G_n}\neg\psi \in cl(\varphi)$ implies $R_{G_1} \cdots R_{G_n}\psi \in cl(\varphi)$;*
6. *$R_{G_1} \cdots R_{G_n}(\psi \wedge \chi) \in cl(\varphi)$ implies $\{R_{G_1} \cdots R_{G_n}\psi, R_{G_1} \cdots R_{G_n}\chi\} \subseteq cl(\varphi)$;*
7. *$R_{G_1} \cdots R_{G_n}K_i\psi \in cl(\varphi)$ implies $D_{\delta(\{i\},G_1,\ldots,G_n)}R_{G_1} \cdots R_{G_n}\psi \in cl(\varphi)$;*
8. *$R_{G_1} \cdots R_{G_n}D_H\psi \in cl(\varphi)$ implies $D_{\delta(H,G_1,\ldots,G_n)}R_{G_1} \cdots R_{G_n}\psi \in cl(\varphi)$;*
9. *$R_{G_1} \cdots R_{G_n}C_H\psi \in cl(\varphi)$ implies all of the following:*
   - *$D_{\delta(H,G_1,\ldots,G_n)}R_{G_1} \cdots R_{G_n}C_H\psi \in cl(\varphi)$,*
   - *$\{D_{\delta(\{i\},G_1,\ldots,G_n)}R_{G_1} \cdots R_{G_n}C_H\psi \mid i \in H\} \subseteq cl(\varphi)$,*
   - *$R_{G_1} \cdots R_{G_n}\psi \in cl(\varphi)$.*

It is not hard to verify that the closure of a formula is finite.

We use $\underline{\Gamma}$ as shorthand for $\bigwedge_{\varphi\in\Gamma} \varphi$ when Γ is a finite set of formulas.

Definition 26 (Canonical pseudo model). *Let α be a formula. The* canonical pseudo model $\mathfrak{M}^c = (S, \backsim, V)$ *for $cl(\alpha)$ is defined below:*

- *$S = \{\Gamma \mid \Gamma$ is maximal consistent in $cl(\alpha)\}$;*
- *$\Gamma \backsim_i \Delta$ iff $\{K_i\varphi \mid K_i\varphi \in \Gamma\} = \{K_i\varphi \mid K_i\varphi \in \Delta\}$;*
- *$\Gamma \backsim_G \Delta$ iff $\{D_H\varphi \mid D_H\varphi \in \Gamma\} = \{D_H\varphi \mid D_H\varphi \in \Delta\}$ whenever $H \subseteq G$;*
- *$V(p) = \{\Gamma \in S \mid p \in \Gamma\}$.*

Proposition 27. *The canonical pseudo model for any $cl(\alpha)$ is a pseudo model.*

Proof. See the appendix. □

Lemma 28. *Let $\mathcal{S} = \{\Gamma \mid \Gamma$ is maximal consistent in $cl(\alpha)\}$ with α a formula. It holds that $\vdash \bigvee_{\Gamma\in\mathcal{S}} \underline{\Gamma}$ and $\vdash \varphi \leftrightarrow \bigvee_{\varphi\in\Gamma\in\mathcal{S}} \underline{\Gamma}$ for all $\varphi \in cl(\alpha)$.*

Proof. See [12, Exercise 7.16] for the first result (although $cl(\alpha)$ is different in our case the proof is exactly the same). We give a proof of the second result in the appendix. □

Let $(S, \backsim|_{G_1}| \cdots |_{G_n}, V)$ be an update of a canonical pseudo model, and $\mathfrak{P} = \langle \Phi_0 \asymp_{\tau_0} \cdots \asymp_{\tau_{n-1}} \Phi_n \rangle$ where $\asymp$ stands for $\backsim|_{G_1}| \cdots |_{G_n}$ and every τ_x is an agent or a group. If all agents in $\tau_0, \ldots, \tau_{n-1}$ appears in H, we call $\mathfrak{P}$ a *$\langle G_1 \cdots G_n\rangle$-resolved H-path (from Φ_0)*; if a formula φ is such that $\varphi \in \Phi_i$ for all $0 \leq i \leq n$, we call $\mathfrak{P}$ a *canonical φ-path*.

Lemma 29. *If Γ and Δ are maximal consistent in $cl(\alpha)$, then*

1. *Γ is* deductively closed *in $cl(\alpha)$, i.e., $\Gamma \vdash \varphi \Leftrightarrow \varphi \in \Gamma$ for any $\varphi \in cl(\alpha)$;*
2. *If $\neg\varphi \in cl(\alpha)$, then $\varphi \in \Gamma \Leftrightarrow \neg\varphi \notin \Gamma$;*
3. *If $\varphi \wedge \psi \in cl(\alpha)$, then $\varphi \wedge \psi \in \Gamma \Leftrightarrow \varphi \in \Gamma$ & $\psi \in \Gamma$;*
4. *If $\underline{\Gamma} \wedge \hat{K}_i\underline{\Delta}$ is consistent, $\Gamma \backsim_i \Delta$; if $\underline{\Gamma} \wedge \hat{D}_G\underline{\Delta}$ is consistent, $\Gamma \backsim_G \Delta$;*
5. *If $K_i\varphi \in cl(\alpha)$, then $K_i\Gamma \vdash \varphi \Leftrightarrow K_i\Gamma \vdash K_i\varphi$;*
6. *If $D_G\varphi \in cl(\alpha)$, then $D_G\Gamma \vdash \varphi \Leftrightarrow D_G\Gamma \vdash D_G\varphi$;*
7. *If $C_G\varphi \in cl(\alpha)$, then $C_G\varphi \in \Gamma \Leftrightarrow \forall\Delta(\Gamma \backsim_{C_G} \Delta \Rightarrow \varphi \in \Delta)$;*
8. *If $R_{G_1} \cdots R_{G_n}C_H\varphi \in cl(\alpha)$, then $R_{G_1} \cdots R_{G_n}C_H\varphi \in \Gamma$ iff every $\langle G_1 \cdots G_n\rangle$-resolved H-path from Γ is a canonical $R_{G_1} \cdots R_{G_n}\varphi$-path.*

PROOF. We give the proof of the clause 8 in the appendix. Other clauses are the same as in [13, Lemma 49] which can be traced back to [12, Chapter 7]. □

LEMMA 30 (PSEUDO TRUTH). *Let $\mathfrak{M}^c = (S, \backsim, V)$ be the canonical pseudo model for $cl(\alpha)$. For all groups $G_1, \ldots, G_n$, all $\Gamma \in S$, and all $R_{G_1} \cdots R_{G_n}\varphi \in cl(\alpha)$, it holds that*

$$R_{G_1} \cdots R_{G_n}\varphi \in \Gamma \quad \textit{iff} \quad \mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Gamma \models \varphi.$$

PROOF. We show this lemma by induction on φ.

- The base case. $R_{G_1} \cdots R_{G_n} p \in \Gamma$ iff $p \in \Gamma$ (Proposition 4(1)) iff $\mathfrak{M}^c, \Gamma \models p$ iff $\mathfrak{M}^c, \Gamma \models R_{G_1} \cdots R_{G_n} p$ iff $\mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Gamma \models p$.
- The case for negation. $R_{G_1} \cdots R_{G_n} \neg\psi \in \Gamma$ iff $\neg R_{G_1} \cdots R_{G_n}\psi \in \Gamma$ (note that $\neg R_{G_1} \cdots R_{G_n}\psi \in cl(\alpha)$ by Definition 25(2,5)) iff $R_{G_1} \cdots R_{G_n}\psi \notin \Gamma$ iff $\mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Gamma \not\models \psi$ iff $\mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Gamma \models \neg\psi$.
- The case for conjunction. $R_{G_1} \cdots R_{G_n}(\psi \wedge \chi) \in \Gamma$ iff $(R_{G_1} \cdots R_{G_n}\psi \wedge R_{G_1} \cdots R_{G_n}\chi) \in \Gamma$ iff $\{R_{G_1} \cdots R_{G_n}\psi, R_{G_1} \cdots R_{G_n}\chi\} \subseteq \Gamma$ ($R_{G_1} \cdots R_{G_n}\psi$ and $R_{G_1} \cdots R_{G_n}\chi$ are in $cl(\alpha)$) iff $\mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Gamma \models \psi$ and $\mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Gamma \models \chi$ iff $\mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Gamma \models \psi \wedge \chi$.
- The case for individual knowledge. From left to right.

$$\begin{array}{lll}
 & R_{G_1} \cdots R_{G_n} K_i\psi \in \Gamma & \\
\text{iff} & D_\delta R_{G_1} \cdots R_{G_n}\psi \in \Gamma \text{ where } \delta = \delta(\{i\}, G_1, \ldots, G_n) & \\
\text{iff} & \forall\Delta.(\Gamma \backsim_\delta \Delta \Rightarrow D_\delta R_{G_1} \cdots R_{G_n}\psi \in \Delta) & \\
\Rightarrow & \forall\Delta.(\Gamma \backsim_\delta \Delta \Rightarrow R_{G_1} \cdots R_{G_n}\psi \in \Delta) & (\mathrm{T}_D) \\
\text{iff} & \forall\Delta.(\Gamma \backsim_\delta \Delta \Rightarrow \mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Delta \models \psi) & (\mathrm{IH}) \\
\text{iff} & \forall\Delta.(\Gamma \backsim_\delta \Delta \Rightarrow \mathfrak{M}^c, \Delta \models R_{G_1} \cdots R_{G_n}\psi) & \\
\text{iff} & \mathfrak{M}^c, \Gamma \models D_\delta R_{G_1} \cdots R_{G_n}\psi & \\
\text{iff} & \mathfrak{M}^c, \Gamma \models R_{G_1} \cdots R_{G_n} K_i\psi & (11(1), 19) \\
\text{iff} & \mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Gamma \models K_i\psi. &
\end{array}$$

  From right to left. Suppose $\mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Gamma \models K_i\psi$. We must show $R_{G_1} \cdots R_{G_n} K_i\psi \in \Gamma$. Suppose this is not the case. Then $\neg R_{G_1} \cdots R_{G_n} K_i\psi \in \Gamma$. Hence $\underline{\Gamma} \wedge \neg R_{G_1} \cdots R_{G_n} K_i\psi$ is consistent, and so is $\underline{\Gamma} \wedge \hat{D}_\delta \neg R_{G_1} \cdots R_{G_n}\psi$, where $\delta = \delta(\{i\}, G_1, \ldots, G_n)$. Let $\mathcal{S}$ be the set of all maximal consistent sets in $cl(\alpha)$. By Lemma 28, $\underline{\Gamma} \wedge \hat{D}_\delta \bigvee_{\neg R_{G_1} \cdots R_{G_n}\psi \in \Theta \in \mathcal{S}} \underline{\Theta}$ is consistent. Since conjunction, resolution and the $\hat{D}_\delta$-operator all distribute over disjunction, $\bigvee_{\neg R_{G_1} \cdots R_{G_n}\psi \in \Theta \in \mathcal{S}} (\underline{\Gamma} \wedge \hat{D}_\delta \underline{\Theta})$ is consistent. Therefore there must be a $\Theta \in \mathcal{S}$ such that $\neg R_{G_1} \cdots R_{G_n}\psi \in \Theta$ and $\underline{\Gamma} \wedge \hat{D}_\delta \underline{\Theta}$ is consistent.

  From $\neg R_{G_1} \cdots R_{G_n}\psi \in \Theta$ we get $R_{G_1} \cdots R_{G_n}\psi \notin \Theta$. By the induction hypothesis $\mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Theta \not\models \psi$, and so $\mathfrak{M}^c, \Theta \not\models R_{G_1} \cdots R_{G_n}\psi$. By Lemma 29(4) and that $\underline{\Gamma} \wedge \hat{D}_\delta \underline{\Theta}$ is consistent, $\Gamma \backsim_\delta \Theta$. But this contradicts the supposition that $\mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Gamma \models K_i\psi$, since by the same reasoning as in the proof of the other direction (see above), $\mathfrak{M}^c, \Delta \models R_{G_1} \cdots R_{G_n}\psi$ for all Δ such that $\Gamma \backsim_\delta \Delta$.
- The case for distributed knowledge: similar to the case for individual knowledge, but just use $\delta(H, G_1, \ldots, G_n)$ instead of $\delta(\{i\}, G_1, \ldots, G_n)$.
- The case for common knowledge. $R_{G_1} \cdots R_{G_n} C_H\psi \in \Gamma$ iff all $\langle G_1 \cdots G_n\rangle$-resolved H-paths from Γ are also canonical $R_{G_1} \cdots R_{G_n}\psi$-paths. Namely, for all Δ such that $(\Gamma, \Delta) \in (\backsim|_{G_1}|\cdots|_{G_n})_{C_H}$, $R_{G_1} \cdots R_{G_n}\psi \in \Delta$ iff for all Δ such that $(\Gamma, \Delta) \in (\backsim|_{G_1}|\cdots|_{G_n})_{C_H}$, it holds by the induction hypothesis that $\mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Delta \models \psi$ iff $\mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Gamma \models C_H\psi$.
- The case for $R_H\psi$. $R_{G_1} \cdots R_{G_n} R_H\psi \in \Gamma$ iff $\mathfrak{M}^c|_{G_1}|\cdots|_{G_n}|_H, \Gamma \models \psi$ (IH applies to ψ) iff $\mathfrak{M}^c|_{G_1}|\cdots|_{G_n}, \Gamma \models R_H\psi$.

□

COROLLARY 31. *Let $\mathfrak{M}^c = (S, \backsim, V)$ be the canonical pseudo model for $cl(\alpha)$. For all $\Gamma \in S$ and all $\varphi \in cl(\alpha)$, it holds that $\varphi \in \Gamma$ iff $\mathfrak{M}^c, \Gamma \models \varphi$.*

LEMMA 32 (PSEUDO COMPLETENESS). *Let φ be an $\mathcal{RCD}$-formula. If φ is valid on all pseudo models, then it is provable in* ***RCD***.

From Pseudo Completeness to Completeness.

By using *unraveling* and *folding* from [13, pp. 9–15], we can transform the canonical pseudo model to a bisimilar pre-model and then to a trans-bisimilar proper model. It remains to show that this process preserves truth. We will use $\rightleftarrows^{\mathrm{T}}$ to denote the trans-bisimulation relation.

LEMMA 33 (INVARIANCE OF TRANS-BISIMULATION). *Let $(\mathfrak{M}, m)$ be a pointed model, $(\mathfrak{N}, n)$ a pointed pre-model, and $(\mathfrak{S}, s)$ a pointed pseudo model. If $(\mathfrak{M}, m) \rightleftarrows^{\mathrm{T}} (\mathfrak{N}, n) \rightleftarrows (\mathfrak{S}, s)$, then $\mathfrak{M}, m \models \varphi$ iff $\mathfrak{N}, n \models_{\mathsf{p}} \varphi$ for all formulas φ.*

PROOF. The lemma can be shown by induction on φ. Here we only show the case for the resolution operators, proofs of other cases are exactly as in the proof of [13, Lemma 26].

Given a pointed model $(\mathfrak{M}, m)$, a pointed pre-model $(\mathfrak{N}, n)$ and a pointed pseudo model $(\mathfrak{S}, s)$, such that $Z : (\mathfrak{M}, m) \rightleftarrows^{\mathrm{T}} (\mathfrak{N}, n)$ for some Z and $(\mathfrak{N}, n) \rightleftarrows (\mathfrak{S}, s)$, we have the following:

$$\begin{array}{rlll}
\mathfrak{M}, m \models R_G\psi & \text{iff} & \mathfrak{M}|_G, m \models \psi & \\
 & \text{iff} & \mathfrak{N}|_G, n \models_{\mathsf{p}} \psi & (*) \\
 & \text{iff} & \mathfrak{N}, n \models_{\mathsf{p}} R_G\psi, &
\end{array}$$

where to show $(*)$ it is sufficient to show that $Z : (\mathfrak{M}|_G, m) \rightleftarrows^{\mathrm{T}} (\mathfrak{N}|_G, n)$, as $(*)$ is then guaranteed by the induction hypothesis (note that $(\mathfrak{N}|_G, n) \rightleftarrows (\mathfrak{S}|_G, s)$ by Proposition 21). Let $\mathfrak{M} = (M, \sim, V)$ and $\mathfrak{N} = (N, \backsim, \nu)$.

- The case for (at) holds by $Z : (\mathfrak{M}, m) \rightleftarrows^{\mathrm{T}} (\mathfrak{N}, n)$.
- As for (zig$_{gr}$), suppose $m(\sim|_G)_H m'$ for some $m' \in M$ and $|H| \geq 2$.
  - If $G \cap H = \emptyset$, $(\sim|_G)_H = \sim_H$. By $Z : (\mathfrak{M}, m) \rightleftarrows^{\mathrm{T}} (\mathfrak{N}, n)$ there is an $n' \in N$ such that $m'Zn'$ and $n \backsim_{H_0} \cdots \backsim_{H_x} n'$ with $H \subseteq H_0 \cap \cdots \cap H_x$. Let $H_0 = \cdots = H_x = H$. Thus $n \backsim_H \cdots \backsim_H n'$. Since $(\backsim|_G)_H = \backsim_H$, it holds that $n(\backsim|_G)_H \cdots (\backsim|_G)_H n'$, and so (zig$_{gr}$) holds in this case.
  - If $G \cap H \neq \emptyset$, $(\sim|_G)_H = \sim_{G\cup H}$. By $Z : (\mathfrak{M}, m) \rightleftarrows^{\mathrm{T}} (\mathfrak{N}, n)$ there is an $n' \in N$ such that $m'Zn'$ and $n \backsim_{H_0} \cdots \backsim_{H_x} n'$ with $G \cup H \subseteq H_0 \cap \cdots \cap H_x$. Thus $n \backsim_{G\cup H} \cdots \backsim_{G\cup H} n'$. Since $(\backsim|_G)_H = \backsim_{G\cup H}$, It holds that $n(\backsim|_G)_H \cdots (\backsim|_G)_H n'$. (zig$_{gr}$) holds also in this case.
- The case for (zig$_{ag}$) is analogous.
- The case for (zag). For all $n' \in N$ and all $\tau \in$ AG ∪ GR, if $n(\backsim|_G)_\tau n'$, then we must show that there is an $m' \in M$ such that $m'Zn'$ and $m(\sim|_G)_\tau m'$.
  - If τ is an agent i. Then if $i \in G$, $(\backsim|_G)_i = \backsim_G$, otherwise $(\backsim|_G)_i = \backsim_i$. By $Z : (\mathfrak{M}, m) \rightleftarrows^{\mathrm{T}} (\mathfrak{N}, n)$, we have $m \sim_G m'$ if $i \in G$, or $m \sim_i m'$ otherwise. Namely $m(\sim|_G)_i m'$ in either case.
  - If τ is a group H. Then if $G \cap H = \emptyset$, $(\backsim|_G)_H = \backsim_H$, otherwise $(\backsim|_G)_H = \backsim_{G\cup H}$. By $Z : (\mathfrak{M}, m) \rightleftarrows^{\mathrm{T}} (\mathfrak{N}, n)$, we have $m \sim_H m'$ if $G \cap H = \emptyset$, or $m \sim_{G\cup H} m'$ otherwise. Namely $m(\sim|_G)_H m'$ in either case.

We have shown that the lemma holds for the case for resolution. For other cases we refer to the proof of [13, Lemma 26]. □

THEOREM 34 (COMPLETENESS). *For any $\mathcal{RCD}$ formula φ, if φ is valid then it is provable in **RCD**.*

PROOF. It suffices to show that any **RCD**-consistent formula is satisfiable. Let φ be consistent. Let $\mathfrak{M}^c$ be the canonical pseudo model for $cl(\varphi)$. By the pseudo truth lemma (with $n = 0$, i.e., an empty list of resolution operators), φ is satisfied in a state Γ in $\mathfrak{M}^c$. Now let $\mathfrak{N}^{\mathfrak{M}^c}$ be the *unraveling* [13, Definition 18][3] of $\mathfrak{M}^c$. $\mathfrak{N}^{\mathfrak{M}^c}$ is a pre-model [13, Proposition 19]. Now let $(\mathfrak{N}^{\mathfrak{M}^c})^*$ be the *folding* [13, Definition 22] of $\mathfrak{N}^{\mathfrak{M}^c}$. $(\mathfrak{N}^{\mathfrak{M}^c})^*$ is a (proper) model [13, Definition 22]. From [13, Lemma 27] and [13, Lemma 28] we have that unraveling preserves bisimulation and that folding preserves trans-bisimulation, in other words we have that[4] $(\mathfrak{M}, \Gamma) \rightleftarrows (\mathfrak{N}^{\mathfrak{M}^c}, \overline{\Gamma}) \rightleftarrows^{\mathsf{T}} ((\mathfrak{N}^{\mathfrak{M}^c})^*, \overline{\Gamma})$. By Corollary 22, $(\mathfrak{N}^{\mathfrak{M}^c}, \overline{\Gamma}) \models_{\mathsf{p}} \varphi$. By Lemma 33, $((\mathfrak{N}^{\mathfrak{M}^c})^*, \overline{\Gamma}) \models \varphi$ and we are done. □

5. DISCUSSION

In this paper we captured the dynamics of publicly observable private resolution of distributed knowledge. Resolution operators (using update semantics) are both an alternative and a complement to the standard distributed knowledge operators (which use standard modal semantics).

Resolution operators let us reason about the relationship between common knowledge and distributed knowledge in general, and in particular about distributed knowledge as potential common knowledge – when can distributed knowledge become common knowledge? A naive idea would be that $D_G\varphi$ should imply that $R_G C_G \varphi$ – any information that is distributed can become common knowledge through resolution. This does not hold in general, however, due to Moore-like phenomena – φ might even become false after resolution (an example is the formula $D_{\{1,2\}}(p \wedge \neg K_1 p)$ discussed in the introduction). We do, however, have the following (Prop. 6(2) with $G = H$):

$$R_G C_G \varphi \leftrightarrow D_G R_G \varphi.$$

A fact can become common knowledge after the group have shared their information if and only if it was distributed knowledge before the event that the fact would be true after the event. This is exactly the distributed knowledge that can become common knowledge (in our special case of publicly observable private resolution of distributed knowledge). If the grand coalition resolves its distributed knowledge, there is no distinction between distributed and common knowledge any more: $R_{\mathrm{AG}} C_{\mathrm{AG}} \varphi \leftrightarrow R_{\mathrm{AG}} D_{\mathrm{AG}} \varphi$ (Prop. 5).

As discussed in the introduction, it has been argued that distributed knowledge in general does not comply with the following *principle of full communication* [11]: if $D_G\varphi$ is true, then φ follows logically from the set of all formulas known by at least one agent in the group. This is seen as a problem: namely that agents can have distributed knowledge without being able to establish it "through communication" [11]. Several papers [11, 4, 7] have tried to characterize classes of models on which the principle of full communication *does* hold – the class of all such models is called *full communication models* [7]. This may seem related to the distinction between distributed knowledge and resolution operators: the latter is intuitively related to internal "full" communication in the group. However, this similarity is superficial: the notion of full communication in the sense of [11] is about *expressive power of the communication language* and the limits that puts on the resulting possible epistemic states under certain assumptions about how information is shared. The key point of the resolution operators, on the other hand, compared with the standard distributed knowledge operators, is to make a distinction between *before* and *after* the information sharing event. That distinction is not made in standard distributed knowledge – even restricted to full communication models: it is easy to see that, e.g., $D_{\{1,2\}}(p \wedge \neg K_1 p)$ is satisfiable also on full communication models. The two ideas, of limiting models to full communication models and of modeling group information sharing events using model updates, are orthogonal, and there is nothing against restricting logics with the resolution operators to full communication models. We leave that for future work. Furthermore, it would be interesting to look at a combined variant: "update by full communication", which takes the communication language into account when defining the updated model.

A main interest for future work is expressive power. Can it be shown that $\mathcal{RCD}$ is strictly more expressive than $\mathcal{ELCD}$? Another, related, natural question is the relative expressivity of $\mathcal{RCD}$ and $\mathcal{PACD}$: can the combination of public announcement operators (which eliminate states) and distributed knowledge operators (which pick out states considered possible by everyone) always be used to "simulate" the resolution operators (which eliminate states considered possible by everyone)?

Also of interest for future work is to look at other assumptions about the other agents' knowledge about the group communication event taking place. In this paper we only studied the case that it is common knowledge that the event takes place (but not what the agents in the group learn). That was naturally modeled using a "global" model update: in every state, replace accessibility for each agent in the group with the group accessibility (intersection). An interesting and also natural alternative is doing only a "local" model update: change accessibility in the same way, but only in the current state. That would correspond to it being common knowledge that if this is the current state, then the group resolves their knowledge.

When looking at the interaction of the resolution and common knowledge operators one might be reminded of *relativized common knowledge* [9, 10]. Here is an open question: can $R_G C_H \varphi$ be expressed using relativized common knowledge, in combination with other operators?

Finally, there is a conceptual relationship to *group announcement logic* [1], where formulas of the form $\langle G \rangle \varphi$ say that G can make a joint public announcement such that φ will become true. A difference to the resolution operators in this paper is that latter model private communication. Yet, the exact relationship between these operators is interesting for future work.

6. ACKNOWLEDGMENTS

Yì N. Wáng acknowledges funding support from the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry of P.R.C.

7. REFERENCES

[1] T. Ågotnes, P. Balbiani, H. van Ditmarsch, and P. Seban. Group announcement logic. *Journal of Applied Logic*, 8(1):62–81, 2010.

[2] A. Baltag, L. Moss, and S. Solecki. The logic of public announcements, common knowledge, and private suspicions. In *Proc. of TARK VII*, pages 43–56, 1998.

[3] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge*. MIT, 1995.

[3] While unraveling is a standard general technique; here we mean unraveling exactly in the sense of the mentioned definition.

[4] Here $\overline{\Gamma}$ is any path in the unraveling starting with Γ.

[4] J. Gerbrandy. *Bisimulations on Planet Kripke*. Ph.D. thesis, University of Amsterdam, 1999.

[5] J.-J. C. Meyer and W. van der Hoek. *Epistemic Logic for AI and Computer Science*. Cambridge University Press, 1995.

[6] J. A. Plaza. Logics of public communications. In *Proceedings of ISMIS*, pages 201–216, 1989.

[7] F. Roelofsen. Distributed knowledge. *Journal of Applied Non-Classical Logics*, 16(2):255–273, 2007.

[8] J. Van Benthem. *Logical dynamics of information and interaction*. Cambridge University Press, 2011.

[9] J. F. A. K. van Benthem. Information update as relativisation. Technical report, ILLC, University of Amsterdam, 2000.

[10] J. F. A. K. van Benthem, J. van Eijck, and B. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

[11] W. van der Hoek, B. van Linder, and J.-J. Meyer. Group knowledge is not always distributed (neither is it always implicit). *Mathematical Social Sciences*, 38(2):215–240, 1999.

[12] H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer, 2007.

[13] Y. Wáng and T. Ågotnes. Public announcement logic with distributed knowledge: Expressivity, completeness and complexity. *Synthese*, 190(1):135–162, 2013.

APPENDIX

Proof of Proposition 6.

1. $\mathfrak{M}, s \models R_G C_H \varphi$ iff $\mathfrak{M}|_G, s \models C_H\varphi$ iff $\mathfrak{M}|_G, t \models \varphi$ for any $(s,t) \in \sim_H^{*'}$, where $\sim_H^{*'} = (\bigcup_{i\in H} \sim_i')^*$ and $\sim_i' = \bigcap_{j\in G} \sim j$ for $i \in G$ and $\sim_i' = \sim_i$ for $i \notin G$. Thus, when $G \cap H = \emptyset$, we get that $\sim_H^{*'} = (\bigcup_{i\in H} \sim_i)^*$. $\mathfrak{M}|_G, t \models \varphi$ for any $(s,t) \in (\bigcup_{i\in H} \sim_i)^*$ holds iff $\mathfrak{M}, t \models R_G\varphi$ for any $(s,t) \in (\bigcup_{i\in H} \sim_i)^*$ iff $\mathfrak{M}, t \models C_H R_G \varphi$.

2.

$$\begin{array}{ll} & \mathfrak{M}, s \models R_G C_H \varphi \\ \text{iff} & \mathfrak{M}|_G, s \models C_H \varphi \\ \text{iff} & \mathfrak{M}|_G, t \models \varphi \text{ for all } t \text{ s.t. } (s,t) \in \sim_{C_H}^G \\ \text{iff}^{\dagger} & \mathfrak{M}|_G, t \models \varphi \text{ for all } t \text{ s.t. } (s,t) \in \sim_i^G \\ \text{iff} & \mathfrak{M}|_G, s \models K_i \varphi \\ \text{iff} & \mathfrak{M}, s \models R_G K_i \varphi \end{array}$$

For the † step, note that when $i \in G$, $\sim_i^G = \sim_j^G$ for any $j \in G$ (and actually also equal to $\sim_G$). Therefore,

$$\sim_{C_H}^G = (\bigcup_{i\in H} \sim_i^G)^* = (\sim_i^G)^* = \sim_i^G .$$

That $R_G K_i \varphi \leftrightarrow D_G R_G \varphi$ is valid is already shown in Proposition 4.

Proof of Proposition 18.

Let $\mathfrak{M} = (S, \backsim, V)$. Clearly $\mathfrak{M}|_G = (S, \backsim|_G, V)$ is a pre-model. Moreover,

1. Given an agent i,

$$\begin{array}{rcl} (\backsim|_G)_{\{i\}} & = & \left\{ \begin{array}{ll} \backsim_{\{i\}\cup G} & i \in G \\ \backsim_{\{i\}}, & i \notin G \end{array} \right. \\ & = & \left\{ \begin{array}{ll} \backsim_G & i \in G \\ \backsim_i, & i \notin G \end{array} \right. \\ & = & (\backsim|_G)_i. \end{array}$$

2. Given two groups H and H' such that $H \subseteq H'$,

$$\begin{array}{rcl} (\backsim|_G)_{H'} & = & \left\{ \begin{array}{ll} \backsim_{H'\cup G} & H' \cap G \neq \emptyset \\ \backsim_{H'}, & H' \cap G = \emptyset \end{array} \right. \\ (\backsim|_G)_H & = & \left\{ \begin{array}{ll} \backsim_{H\cup G} & H \cap G \neq \emptyset \\ \backsim_H, & H \cap G = \emptyset \end{array} \right. \end{array}$$

So we have:

- when $H \cap G \neq \emptyset$ (and therefore $H' \cap G \neq \emptyset$), $(\backsim|_G)_{H'} = \backsim_{H'\cup G} \subseteq \backsim_{H\cup G} = (\backsim|_G)_H$;
- when $H' \cap G = \emptyset$ (and therefore $H \cap G = \emptyset$), $(\backsim|_G)_{H'} = \backsim_{H'} \subseteq \backsim_H = (\backsim|_G)_H$;
- otherwise $H' \cap G \neq \emptyset$ and $H \cap G = \emptyset$, and in this case $(\backsim|_G)_{H'} = \backsim_{H'\cup G} \subseteq \backsim_H = (\backsim|_G)_H$.

$\mathfrak{M}|_G$ is a pre-model satisfying the two conditions above, which shows it is a pseudo model.

Proof of Proposition 21.

Let $\mathfrak{M} = (S, \backsim, V)$ and $\mathfrak{M}' = (S', \backsim', V')$. Thus $\mathfrak{M}|_G = (S, \backsim|_G, V)$ and $\mathfrak{M}'|_G = (S', \backsim'|_G, V')$. Suppose $Z : (\mathfrak{M}, s) \rightleftarrows (\mathfrak{M}', s')$, and we show $Z : (\mathfrak{M}|_G, s) \rightleftarrows (\mathfrak{M}'|_G, s')$:

(at) This clearly follows from the (at) clause of $Z : (\mathfrak{M}, s) \rightleftarrows (\mathfrak{M}', s')$.

(zig) For all $t \in S$, if $s(\backsim|_G)_H t$, then

- If $G \cap H = \emptyset$, then $(\backsim|_G)_H = \backsim_H$ and $(\backsim'|_G)_H = \backsim'_H$. By $Z : (\mathfrak{M}, s) \rightleftarrows (\mathfrak{M}', s')$ there must be a $t' \in S'$ such that $s'(\backsim'|_G)_H t'$ and tZt'.
- If $G \cap H \neq \emptyset$, then $(\backsim|_G)_H = \backsim_{G\cup H}$ and $(\backsim'|_G)_H = \backsim'_{G\cup H}$. By $Z : (\mathfrak{M}, s) \rightleftarrows (\mathfrak{M}', s')$ there must be a $t' \in S'$ such that $s'(\backsim'|_G)_H t'$ and tZt'.

If $s(\backsim|_G)_i t$, we can prove analogously to the above.

(zag) This can be shown analogously to the case for (zig).

Proof of Theorem 24.

It is easy to verify that (**S5**), (**CK**), (**DK**), (N$_R$), (RA), (RC) and (RN) are all valid or admissible with respect to the class of all pseudo models. Here we only show that i) (RD1) and (RD2) are valid in all pseudo models, and ii) (RR$_C$) preserves validity of pseudo models.

Let $\mathfrak{M} = (S, \backsim, V)$ be a pseudo model and $s \in S$. We show the following:

- $\mathfrak{M}, s \models_{\mathsf{p}}$ RD1 and $\mathfrak{M}, s \models_{\mathsf{p}}$ RD2, i.e.,
  - If $G \cap H \neq \emptyset$, then $\mathfrak{M}, s \models_{\mathsf{p}} R_G D_H \varphi \leftrightarrow D_{G\cup H} R_G \varphi$;
  - If $G \cap H = \emptyset$, then $\mathfrak{M}, s \models_{\mathsf{p}} R_G D_H \varphi \leftrightarrow D_H R_G \varphi$.

  $$\begin{array}{ll} & \mathfrak{M}, s \models_{\mathsf{p}} R_G D_H \varphi \\ \text{iff} & \mathfrak{M}|_G, s \models_{\mathsf{p}} D_H \varphi \\ \text{iff} & \mathfrak{M}|_G, t \models_{\mathsf{p}} \varphi \text{ for all } t \text{ s.t. } (s,t) \in (\backsim|_G)_H \\ \text{iff} & \mathfrak{M}, t \models_{\mathsf{p}} R_G\varphi \text{ for all } t \text{ s.t. } (s,t) \in (\backsim|_G)_H \\ \text{iff}^{\dagger} & \text{if } G \cap H \neq \emptyset,\ \mathfrak{M}, t \models_{\mathsf{p}} R_G\varphi \text{ for all } t \text{ s.t. } (s,t) \in \backsim|_{G\cup H}, \\ & \text{if } G \cap H = \emptyset,\ \mathfrak{M}, t \models_{\mathsf{p}} R_G\varphi \text{ for all } t \text{ s.t. } (s,t) \in \backsim|_H \\ \text{iff} & \text{if } G \cap H \neq \emptyset,\ \mathfrak{M}, t \models_{\mathsf{p}} D_{G\cup H} R_G \varphi, \text{ and} \\ & \text{if } G \cap H = \emptyset,\ \mathfrak{M}, t \models_{\mathsf{p}} D_H R_G \varphi, \end{array}$$

  where the † step is by definition:

  $$(\backsim|_G)_H = \left\{ \begin{array}{ll} \backsim_{H\cup G}, & G \cap H \neq \emptyset, \\ \backsim_H, & G \cap H = \emptyset. \end{array} \right.$$

- $\mathfrak{M}, s \models_{\mathsf{p}} \varphi \to R_{G_1} \cdots R_{G_n} C_H \psi$ under the assumption $\models_{\mathsf{p}} \varphi \to (E_H \varphi \wedge R_{G_1} \cdots R_{G_n} \psi)$. The proof is similar to the proof for genuine models.

Proof of Proposition 27.

Suppose that $\mathfrak{M} = (S, \backsim, V)$ is the canonical pseudo model for $cl(\alpha)$. We need to show that $\mathfrak{M}$ is a pseudo model. Namely,

1. S is non-empty, and
2. all $\backsim_i$'s and $\backsim_G$'s are equivalence relations, and
3. V is a valuation from PROP to $\wp(S)$, and
4. $\backsim_i = \backsim_{\{i\}}$ for every agent i, and
5. $\backsim_H \subseteq \backsim_G$ if G and H are groups such that $G \subseteq H$.

Conditions 1–3 are the conditions for being a pre-model which are easy to verify. Conditions 4 and 5 are additional conditions for being a pseudo model.

By Definition 25(3), $K_i\varphi$ and $D_i\varphi$ must be in $cl(\alpha)$ both or neither. Thus, for any $\Gamma, \Delta \in S$,

$$\begin{array}{ll} & \Gamma \backsim_i \Delta \\ \text{iff} & \{K_i\varphi \mid K_i\varphi \in \Gamma\} = \{K_i\varphi \mid K_i\varphi \in \Delta\} \\ \text{iff} & \{D_i\varphi \mid D_i\varphi \in \Gamma\} = \{D_i\varphi \mid D_i\varphi \in \Delta\} \quad \text{(Axiom DK1)} \\ \text{iff} & \Gamma \backsim_{\{i\}} \Delta. \end{array}$$

$$\begin{array}{ll} & \Gamma \backsim_H \Delta \\ \text{iff} & \{D_{H'}\varphi \mid D_{H'}\varphi \in \Gamma\} = \{D_{H'}\varphi \mid D_{H'}\varphi \in \Delta\}, \\ & \text{for any group } H' \subseteq H \\ \Rightarrow & \{D_{G'}\varphi \mid D_{G'}\varphi \in \Gamma\} = \{D_{G'}\varphi \mid D_{G'}\varphi \in \Delta\}, \\ & \text{for any group } G' \subseteq G \\ \text{iff} & \Gamma \backsim_G \Delta. \end{array}$$

This finishes the proof, and shows that the notion "canonical pseudo model" is well-defined.

Proof of Lemma 28(2).

Let $\varphi \in cl(\alpha)$. By $\vdash \bigvee_{\neg\varphi\in\Gamma\in\mathcal{S}} \underline{\Gamma} \rightarrow \neg\varphi$ and the first result (i.e., $\vdash \bigvee_{\neg\varphi\in\Gamma\in\mathcal{S}} \vee \bigvee_{\varphi\in\Gamma\in\mathcal{S}}$) we get $\vdash \varphi \rightarrow \bigvee_{\varphi\in\Gamma\in\mathcal{S}} \underline{\Gamma}$. For the converse direction, suppose $\nvdash \bigvee_{\varphi\in\Gamma\in\mathcal{S}} \underline{\Gamma} \rightarrow \varphi$. Then $\neg(\bigvee_{\varphi\in\Gamma\in\mathcal{S}} \underline{\Gamma} \rightarrow \varphi)$ is consistent. Namely $\neg\varphi \wedge \bigvee_{\varphi\in\Gamma\in\mathcal{S}} \underline{\Gamma}$ is consistent. But this is impossible.

Proof of Lemma 29(8).

Let $R_{G_1} \cdots R_{G_n} C_H\varphi \in cl(\alpha)$. It follows from the definition of closure (Definition 25) that the following formulas:

- $D_{\delta(\{i\},G_1,\ldots,G_n)} R_{G_1} \cdots R_{G_n} C_H\varphi$ where $i \in H$
- $D_{\delta(H,G_1,\ldots,G_n)} R_{G_1} \cdots R_{G_n} C_H\varphi$
- $R_{G_1} \cdots R_{G_n}\varphi$ and $\neg R_{G_1} \cdots R_{G_n}\varphi$

are all in $cl(\alpha)$.

From left to right. Suppose $R_{G_1} \cdots R_{G_n} C_H\varphi \in \Gamma$, we continue by induction on the length of the path that every $\langle G_1 \cdots G_n\rangle$-resolved H-path from Γ is a canonical $R_{G_1} \cdots R_{G_n} C_H\varphi$-path. Then the left-to-right direction follows: by $\vdash C_H\varphi \rightarrow \varphi$, N$_R$ and R_G-distribution (which follows from **RR** axioms) we get $\vdash R_{G_1} \cdots R_{G_n} C_H\varphi \rightarrow R_{G_1} \cdots R_{G_n}\varphi$, and by $R_{G_1} \cdots R_{G_n}\varphi \in cl(\alpha)$ we have $R_{G_1} \cdots R_{G_n}\varphi \in \Gamma$.

Suppose the length of the $\langle G_1 \cdots G_n\rangle$-resolved H-path is 0, i.e., the path is $\langle\Gamma\rangle$, we must show that $R_{G_1} \cdots R_{G_n} C_H\varphi \in \Gamma$. This is guaranteed by the supposition.

Suppose the length of the $\langle G_1 \cdots G_n\rangle$-resolved H-path is $n+1$, i.e., the path is $\langle\Gamma_0 \asymp_{\tau_0} \cdots \asymp_{\tau_{n-1}} \Gamma_n \asymp_{\tau_n} \Gamma_{n+1}\rangle$ with $\Gamma_0 = \Gamma$ and every τ_x is either in H or a subset of H. By the induction hypothesis we may assume that $R_{G_1} \cdots R_{G_n} C_H\varphi \in \Gamma_n$.

- Suppose τ_n is an agent i ($i \in H$). By Axiom C1 we have $\vdash C_H\varphi \rightarrow K_i C_H\varphi$. It follows that $\vdash R_{G_1} \cdots R_{G_n} C_H\varphi \rightarrow R_{G_1} \cdots R_{G_n} K_i C_H\varphi$ by the rules N$_R$ and R_G-distribution. Let $\delta = \delta(\{i\}, G_1, \ldots, G_n)$. By the reduction axioms we move K_i left, i.e., $\vdash R_{G_1} \cdots R_{G_n} K_i C_H\varphi \rightarrow D_\delta R_{G_1} \cdots R_{G_n} C_H\varphi$, so we get $\vdash R_{G_1} \cdots R_{G_n} C_H\varphi \rightarrow D_\delta R_{G_1} \cdots R_{G_n} C_H\varphi$. Hence $\Gamma_n \vdash D_\delta R_{G_1} \cdots R_{G_n} C_H\varphi$. As $D_\delta R_{G_1} \cdots R_{G_n} C_H\varphi \in cl(\alpha)$, we have $D_\delta R_{G_1} \cdots R_{G_n} C_H\varphi \in \Gamma_n$. Moreover, by Proposition 19, $\asymp_i = \backsim_\delta$. Thus $D_\delta R_{G_1} \cdots R_{G_n} C_H\varphi \in \Gamma_{n+1}$ by the definition of $\backsim_\delta$, and so $R_{G_1} \cdots R_{G_n} C_H\varphi \in \Gamma_{n+1}$.
- Suppose τ_n is a group I ($I \subseteq H$). By Axioms C1, D1 and D2 we have $\vdash C_H\varphi \rightarrow D_I C_H\varphi$. By N$_R$ and R_G-distribution, $\vdash R_{G_1} \cdots R_{G_n} C_H\varphi \rightarrow R_{G_1} \cdots R_{G_n} D_I C_H\varphi$. By similar reasoning to the case above, we get the result $R_{G_1} \cdots R_{G_n} C_H\varphi \in \Gamma_{n+1}$ (we use $\delta(H, G_1, \ldots, G_n)$ instead of $\delta(\{i\}, G_1, \ldots, G_n)$ in this case).

In both cases we get $R_{G_1} \cdots R_{G_n} C_H\varphi \in \Gamma_{n+1}$ as we wish to show.

From right to left. Suppose that every $\langle G_1 \cdots G_n\rangle$-resolved H-path from Γ is a canonical $R_{G_1} \cdots R_{G_n}\varphi$-path. Let $\mathcal{S}_0$ be the set of all maximal consistent sets Δ in $cl(\alpha)$ such that every $\langle G_1 \cdots G_n\rangle$-resolved H-path from Δ is a canonical $R_{G_1} \cdots R_{G_n}\varphi$-path. Now consider the formula

$$\lambda = \bigvee_{\Delta\in\mathcal{S}_0} \underline{\Delta}$$

We will show the following:

1. $\vdash \underline{\Gamma} \rightarrow \lambda$
2. $\vdash \lambda \rightarrow (E_H\lambda \wedge R_{G_1} \cdots R_{G_n}\varphi)$.

From the above and the reduction rule for resolved common knowledge we get $\vdash \underline{\Gamma} \rightarrow R_{G_1} \cdots R_{G_n} C_H\varphi$ which furthermore entails $R_{G_1} \cdots R_{G_n} C_H\varphi \in \Gamma$. We now continue with the proof of the two clauses.

1. This is trivial, as $\underline{\Gamma}$ is one of the disjuncts of λ.
2. Suppose towards a contradiction that

$$\lambda \wedge \neg(E_H\lambda \wedge R_{G_1} \cdots R_{G_n}\varphi)$$

is consistent, i.e., $\lambda \wedge (\neg E_H\lambda \vee \neg R_{G_1} \cdots R_{G_n}\varphi)$ is consistent. Because λ is a disjunction there must be a disjunct $\underline{\Xi}$ of λ such that $\underline{\Xi} \wedge (\neg E_H\lambda \vee \neg R_{G_1} \cdots R_{G_n}\varphi)$ is consistent. It follows that either $\underline{\Xi} \wedge \neg E_H\lambda$ or $\underline{\Xi} \wedge \neg R_{G_1} \cdots R_{G_n}\varphi$ is consistent.
If the former is consistent, then there must be an agent $i \in H$ such that $\underline{\Xi} \wedge \neg K_i\lambda$ is consistent, i.e., $\underline{\Xi} \wedge \hat{K}_i \neg \bigvee_{\Delta\in\mathcal{S}_0} \underline{\Delta}$ is consistent. Since $\vdash \bigvee_{\Delta\in\mathcal{S}} \underline{\Delta}$ by Lemma 28, we have $\vdash \neg \bigvee_{\Delta\in\mathcal{S}_0} \underline{\Delta} \rightarrow \bigvee_{\Delta'\in\mathcal{S}\setminus\mathcal{S}_0} \underline{\Delta'}$, and so there must be a Θ in $\mathcal{S} \setminus \mathcal{S}_0$ such that $\underline{\Xi} \wedge \hat{K}_i\underline{\Theta}$ is consistent. By item 4 of this lemma $\Xi \backsim_i \Theta$ (where $\backsim$ is the relation in the canonical pseudo model for $cl(\alpha)$). But then Ξ cannot be in $\mathcal{S}_0$ for $\Theta \notin \mathcal{S}_0$. A contradiction!
If the latter is consistent, since $\neg R_{G_1} \cdots R_{G_n}\varphi \in cl(\alpha)$ and Ξ is maximal, $\neg R_{G_1} \cdots R_{G_n}\varphi \in \Xi$. But $R_{G_1} \cdots R_{G_n}\varphi \in \Xi$ since $\Xi \backsim_H \Xi$ and every $\langle G_1 \cdots G_n\rangle$-resolved H-path from Ξ is a canonical $R_{G_1} \cdots R_{G_n}\varphi$-path. We reach a contradiction.

Epistemic Protocols for Distributed Gossiping

Krzysztof R. Apt
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
k.r.apt@cwi.nl

Davide Grossi
University of Liverpool
Liverpool, UK
d.grossi@liverpool.ac.uk

Wiebe van der Hoek
University of Liverpool
Liverpool, UK
wiebe@liv.ac.uk

ABSTRACT

Gossip protocols aim at arriving, by means of point-to-point or group communications, at a situation in which all the agents know each other's secrets. We consider distributed gossip protocols which are expressed by means of epistemic logic. We provide an operational semantics of such protocols and set up an appropriate framework to argue about their correctness. Then we analyze specific protocols for complete graphs and for directed rings.

Keywords

Gossip protocols, epistemic logic, distributed computing, knowledge-based programs

1. INTRODUCTION

In the gossip problem ([18, 4], see also [10] for an overview) a number n of agents, each one knowing a piece of information (a *secret*) unknown to the others, communicate by one-to-one interactions (e.g., telephone calls). The result of each call is that the two agents involved in it learn all secrets the other agent knows at the time of the call. The problem consists in finding a sequence of calls which disseminates all the secrets among the agents in the group. It sparked a large literature in the 70s and 80s [18, 4, 9, 5, 17] typically focusing on establishing—in the above and other variants of the problem—the minimum number of calls to achieve dissemination of all the secrets. This number has been proven to be $2n-4$, where n, the number of agents, is at least 4.

The above literature assumes a centralized perspective on the gossip problem: a planner schedules agents' calls. In this paper we pursue a line of research first put forth in [3] by developing a decentralized theory of the gossip problem, where agents perform calls not according to a centralized schedule, but following individual epistemic protocols they run in a distributed fashion. These protocols tell the agents which calls to execute depending on what they know, or do not know, about the information state of the agents in the group. We call the resulting distributed programs *(epistemic) gossip protocols.*

Contribution of the paper and outline.

The paper introduces a formal framework for specifying epistemic gossip protocols and for studying their computations in terms of correctness, termination, and fair termination (Section 2). It then defines and studies two natural protocols in which the interactions are unconstrained (Section 3) and four example gossip protocols in which agents are positioned on a directed ring and calls can happen only between neighbours (Section 4). Proofs are collected in the appendix.

From a methodological point of view, the paper integrates concepts and techniques from the distributed computing, see, e.g., [1, Chapter 11] and the epistemic logic literature [8, 15] in the tradition of [16, 14, 7].

2. GOSSIP PROTOCOLS

We introduce first the syntax and semantics of gossip protocols.

2.1 Syntax

We loosely use the syntax of the language CSP (Communicating Sequential Processes) of [11] that extends the guarded command language of [6] by disjoint parallel composition and commands for synchronous communication. CSP was realized in the distributed programming language OCCAM (see INMOS [12]).

The main difference is that we use as guards epistemic formulas and as communication primitives calls that do not require synchronization. Also, the syntax of our distributed programs is very limited. In order to define gossip protocols we introduce in turn calls and epistemic guards.

Throughout the paper we assume a fixed finite set A of at least three ***agents***. We assume that each agent holds exactly one ***secret*** and that there exists a bijection between the set of agents and the set of secrets. We denote by P the set of all secrets (for *p*ropositions). Furthermore, it is assumed that each secret carries information identifying the agent to whom that secret belongs.

2.1.1 Calls

Each ***call*** concerns two agents, the *caller* (a below) and the *agent called* (b). We distinguish three ***modes of communication*** of a call:

push-pull, written as ab or (a, b). During this call the caller and the called agent learn each other's secrets,

push, written as $a \rhd b$. After this call the called agent learns all the secrets held by the caller,

pull, written as $a \lhd b$. After this call the caller learns all the secrets held by the called agent.

Variables for calls are denoted by c, d. Abusing notation we write $a \in \mathsf{c}$ to denote that agent a is one of the two agents involved in the call c (e.g., for $\mathsf{c} := ab$ we have $a \in \mathsf{c}$ and $b \in \mathsf{c}$). Calls in which agent a is involved are denoted by c^a.

2.1.2 Epistemic guards

Epistemic guards are defined as formulas in a simple modal language with the following grammar:

$$\phi ::= F_a p \mid \neg\phi \mid \phi \wedge \phi \mid K_a\phi,$$

where $p \in \mathsf{P}$ and $a \in \mathsf{A}$. Each secret is viewed as a distinct symbol. We denote the secret of agent a by A, the secret of agent b by B and so on. We denote the set of so defined formulas by $\mathcal{L}$ and we refer to its members as epistemic formulas or epistemic guards. We read $F_a p$ as 'agent a is familiar with the secret p' (or 'p belongs to the set of secrets a knows about') and $K_a\phi$ as 'agent a knows that formula ϕ is true'. So this language is an epistemic language where atoms consist of 'knowing whether' statements about propositional atoms, if we view secrets as Boolean variables.

Atomic expressions in $\mathcal{L}$ concern only who knows what secrets. As a consequence the language cannot express formally the truth of a secret p. This level of abstraction suffices for the purposes of the current paper. However, expressions $F_a p$ could be given a more explicit epistemic reading in terms of 'knowing whether'. That is, 'a is familiar with p' can be interpreted (on a suitable Kripke model) as 'a knows whether the secret p is true or not'. This link is established in [3].

2.1.3 Gossip protocols

Before specifying what a program for agent a is, let us first define the language $\mathcal{L}_a$ with the following grammar:

$$\psi ::= K_a\phi \mid \neg\psi \mid \psi \wedge \psi$$

with $\phi \in \mathcal{L}$.[1]

By a ***component program***, in short a ***program***, for an agent a we mean a statement of the form

$$*[[]_{j=1}^{m} \ \psi_j \to \mathsf{c}_j],$$

where $m > 0$ and each $\psi_j \to \mathsf{c}_j$ is such that $\psi_j \in \mathcal{L}_a$ and a is the caller in c_j.

Given an epistemic formula $\psi \in \mathcal{L}_a$ and a call c, we call the construct $\psi \to \mathsf{c}$ a ***rule*** and refer in this context to ψ as a ***guard***.

We denote the set of rules $\{\psi_1 \to \mathsf{c}_1, \ldots, \psi_k \to \mathsf{c}_k\}$ as $[[]_{j=1}^{k} \ \psi_j \to \mathsf{c}_j]$ and abbreviate a set of rules $\{\psi_1 \to \mathsf{c}, \ldots, \psi_k \to \mathsf{c}\}$ with the same call to a single rule $\bigvee_{i=1}^{k} \psi_i \to \mathsf{c}$.

Intuitively, $*$ denotes a repeated execution of the rules, one at a time, where each time a rule is selected whose guard is true.

Finally, by a ***distributed epistemic gossip protocol***, in short a ***gossip protocol***, we mean a parallel composition of component programs, one for each agent. In order not to complicate matters we assume that each gossip protocol uses only one mode of communication.

Of special interest for this paper are gossip protocols that are symmetric. By this we mean that the protocol is a composition of the component programs that are identical modulo the names of the agents. Formally, consider a statement $\pi(x)$, where x is a variable ranging over the set A of agents and such that for each agent $a \in \mathsf{A}$, $\pi(a)$ is a component program for agent a. Then the parallel composition of the $\pi(a)$ programs, where $a \in \mathsf{A}$, is called a ***symmetric gossip protocol***.

Gossip protocols are syntactically extremely simple. Therefore it would seem that little can be expressed using them. However, this is not the case. In Sections 3 and 4 we consider gossip protocols that can exhibit complex behaviour.

2.2 Semantics

We now move on to provide a formal semantics of epistemic guards, and then describe the computations of gossip protocols.

2.2.1 Gossip situations and calls

A ***gossip situation*** is a sequence $\mathsf{s} = (\mathsf{Q}_a)_{a \in \mathsf{A}}$, where $\mathsf{Q}_a \subseteq \mathsf{P}$ for each agent a. Intuitively, Q_a is the set of secrets a is familiar with in situation s. The ***initial gossip situation*** is the one in which each Q_a equals $\{A\}$ and is denoted by root. The set of all gossip situations is denoted by S. We say that an agent a is an ***expert*** in a gossip situation s if he is familiar in s with all the secrets, i.e., if $\mathsf{Q}_a = \mathsf{P}$. The initial gossip situation reflects the fact that initially each agent is familiar only with his own secret, although it is not assumed this is common knowledge among the agents. In fact, in the introduced language we have no means to express the concept of common knowledge.

We will use the following concise notation for gossip situations. Sets of secrets will be written down as lists. e.g., the set $\{A, B, C\}$ will be written as ABC. Gossip situations will be written down as lists of lists of secrets separated by dots. E.g., if there are three agents, $\mathsf{root} = A.B.C$ and the situation $(\{A, B\}, \{A, B\}, \{C\})$ will be written as $AB.AB.C$.

Each call transforms the current gossip situation by modifying the set of secrets the agents involved in the call are familiar with. More precisely, the application of a call to a situation is defined as follows.

Definition 2.1 (Effects of calls). *A call is a function $\mathsf{c} : \mathsf{S} \longrightarrow \mathsf{S}$, so defined, for $\mathsf{s} := (\mathsf{Q}_a)_{a \in \mathsf{A}}$:*

$\boxed{\mathsf{c} = ab}$ $\mathsf{c}(\mathsf{s}) = (Q'_a)_{a \in \mathsf{A}}$, *where* $\mathsf{Q}'_a = \mathsf{Q}'_b = \mathsf{Q}_a \cup \mathsf{Q}_b$, $\mathsf{Q}'_c = \mathsf{Q}_c$, *for* $c \neq a, b$;

$\boxed{\mathsf{c} = a \rhd b}$ $\mathsf{c}(\mathsf{s}) = (Q'_a)_{a \in \mathsf{A}}$, *where* $\mathsf{Q}'_b = \mathsf{Q}_a \cup \mathsf{Q}_b$, $\mathsf{Q}'_a = \mathsf{Q}_a$, $\mathsf{Q}'_c = \mathsf{Q}_c$, *for* $c \neq a, b$;

$\boxed{\mathsf{c} = a \lhd b}$ $\mathsf{c}(\mathsf{s}) = (Q'_a)_{a \in \mathsf{A}}$, *where* $\mathsf{Q}'_a = \mathsf{Q}_a \cup \mathsf{Q}_b$, $\mathsf{Q}'_b = \mathsf{Q}_b$, $\mathsf{Q}'_c = \mathsf{Q}_c$, *for* $c \neq a, b$.

The definition formalizes the modes of communications we introduced earlier. Depending on the mode, secrets are either shared between caller and callee (ab), they are pushed from the caller to the callee ($a \rhd b$), or they are retrieved by the caller from the callee ($a \lhd b$).

2.2.2 Call sequences

A ***call sequence*** is a (possibly infinite) sequence of calls, in symbols $(\mathsf{c}_1, \mathsf{c}_2, \ldots, \mathsf{c}_n, \ldots)$, all being of the same communication mode. The empty sequence is denoted by ϵ. We use $\mathbf{c}$ to denote a call sequence and $\mathbf{C}$ to denote the set of all call sequences. The set of all finite call sequences is denoted $\mathbf{C}^{<\omega}$. Given a finite call sequence $\mathbf{c}$ and a call c we denote by $\mathsf{c}.\mathbf{c}$ the prepending of $\mathbf{c}$ with c, and by $\mathbf{c}.\mathsf{c}$ the postpending of $\mathbf{c}$ with c.

[1] Alternatively, $\mathcal{L}_a$ could be defined as the fragment of $\mathcal{L}$ consisting of the formulae of form $K_a\psi$. In logic S5, it is easy to prove that each $\psi \in \mathcal{L}_a$ is logically equivalent to a formula $K_a\phi \in \mathcal{L}$.

The result of applying a call sequence to a situation s is defined by induction using Definition 2.1, as follows:
[Base] $\epsilon(\mathsf{s}) := \mathsf{s}$,
[Step] $(\mathsf{c}.\mathbf{c})(\mathsf{s}) := \mathbf{c}(\mathsf{c}(\mathsf{s}))$.

Example 2.2. *Let the set of agents be $\{a, b, c\}$.*

$$\begin{array}{ccccccc} & ab & & ca & & ab & \\ A.B.C & & AB.AB.C & & ABC.AB.ABC & & ABC.ABC.ABC \end{array}$$

The top row lists the call sequence (ab, ca, ab), while the bottom row lists the successive gossip situations obtained from the initial situation $A.B.C$ by applying the calls in the sequence: first ab, then ca and finally ab. □

By applying an infinite call sequence $\mathbf{c} = (\mathsf{c}_1, \mathsf{c}_2, \ldots, \mathsf{c}_n, \ldots)$ to a gossip situation s one obtains therefore an infinite sequence $\mathbf{c}^0(\mathsf{s}), \mathbf{c}^1(\mathsf{s}), \ldots, \mathbf{c}^n(\mathsf{s}), \ldots$ of gossip situations, where each $\mathbf{c}^k$ is sequence $\mathsf{c}_1, \mathsf{c}_2, \ldots, \mathsf{c}_k$. A call sequence $\mathbf{c}$ is said to ***converge*** if for all input gossip situations s the generated sequence of gossip situations reaches a limit, that is, there exists $n < \omega$ such that for all $m \geq n$ $\mathbf{c}^m(\mathsf{s}) = \mathbf{c}^{m+1}(\mathsf{s})$. Since the set of secrets is finite and calls never make agents forget secrets they are familiar with, it is easy to see the following.

Fact 2.3. *All infinite call sequences converge.*

However, as we shall see, this does not imply that all gossip protocols terminate. In the remainder of the paper, unless stated otherwise, we will assume the push-pull mode of communication. The reader can easily adapt our presentation to the other modes.

2.2.3 Gossip models

The set S of all gossip situations is the set of all possible combinations of secret distributions among the agents. As calls progress in sequence from the initial situation, agents may be uncertain about which one of such secrets distributions is the actual one. This uncertainty is precisely the object of the epistemic language for guards we introduced earlier.

Definition 2.4. *A* **gossip model** *(for a given set A) is a tuple $\mathcal{M} = (\mathbf{C}^{<\omega}, \{\sim_a\}_{a \in \mathsf{A}})$, where each $\sim_a \subseteq \mathbf{C}^{<\omega} \times \mathbf{C}^{<\omega}$ is the smallest relation satisfying the following inductive conditions (assume the mode of communication is push-pull):*

[Base] $\epsilon \sim_a \epsilon$;

[Step] *Suppose* $\mathbf{c} \sim_a \mathbf{d}$.

(i) If $a \notin \mathsf{c}$, then $\mathbf{c}.\mathsf{c} \sim_a \mathbf{d}$ and $\mathbf{c} \sim_a \mathbf{d}.\mathsf{c}$.

(ii) If there exists $b \in \mathsf{A}$ and $\mathsf{c}, \mathsf{d} \in \{ab, ba\}$ such that $\mathbf{c}.\mathsf{c}(\mathsf{root})_a = \mathbf{d}.\mathsf{d}(\mathsf{root})_a$, then $\mathbf{c}.\mathsf{c} \sim_a \mathbf{d}.\mathsf{d}$.

A gossip model with a designated finite call sequence is called a **pointed gossip model**.

For the push, respectively pull, modes of communication clause (ii) needs to be modified by requiring that for some $b \in \mathsf{A}$, $\mathsf{c} = \mathsf{d} = a \rhd b$ or $\mathsf{c} = \mathsf{d} = a \lhd b$, respectively.

For instance, by *(i)* we have $ab, bc \sim_a ab, bd$. But we do not have $bc, ab \sim_a bd, ab$ since $(bc, ab)(\mathsf{root})_a = ABC \neq ABD = (bd, ab)(\mathsf{root})_a$.

Let us flesh out the intuitions behind the above definition. Gossip models are needed in order to interpret the epistemic guards of gossip protocols. Since such guards are relevant only after finite sequences of calls, the domain of a gossip model is taken to consist only of finite sequences. Intuitively, those are the finite sequences that can be generated by a gossip protocol. Let us turn now to the $\sim_a$ relation. This is defined with the following intuitions in mind. First of all, no agent can distinguish the empty call sequence from itself—this is the base of the induction. Next, if two call sequences are indistinguishable for a, then the same is the case if *(i)* we extend one of these sequences by a call in which a is not involved or if *(ii)* we extend each of these sequences by a call of a with the same agent (agent a may be the caller or the callee), provided a is familiar with exactly the same secrets after each of the new sequences has taken place—this is the induction step.[2]

The above intuitions are based on the following assumptions on the form of communication we presuppose: (i) At the initial situation, as communication starts, each agent knows only her own secret but considers it possible that the others may be familiar with all other secrets. In other words there is no such thing as common knowledge of the fact that 'everybody knows exactly her own secret'. (ii) In general, each agent always considers it possible that call sequences (of any length) take place that do not involve her. These assumptions are weaker than the ones analyzed in [3].

We state without proof the following simple fact.

Fact 2.5.

(i) Each $\sim_a$ is an equivalence relation;

(ii) For all $\mathbf{c}, \mathbf{d} \in \mathbf{C}$ if $\mathbf{c} \sim_a \mathbf{d}$, then $\mathbf{c}(\mathsf{root})_a = \mathbf{d}(\mathsf{root})_a$, but not vice versa.

This prompts us to note also that according to Definition 2.4 sequences which make a learn the same set of secrets may well be distinguishable for a, such as, for instance, ab, bc, ab and ab, bc, ac. In the first one a comes to know that b knows a is familiar with all secrets, while in the second one, she comes to know that c knows a is familiar with all secrets. Relation $\sim_a$ is so defined as to capture this sort of 'higher-order' knowledge.

2.2.4 Truth conditions for epistemic guards

Everything is now in place to define the truth of the considered formulas.

Definition 2.6. *Let $(\mathcal{M}, \mathbf{c})$ be a pointed gossip model with $\mathcal{M} = (\mathbf{C}^{<\omega}, (\sim_a)_{a \in \mathsf{A}})$ and $\mathbf{c} \in \mathbf{C}^{<\omega}$. We define the satisfaction relation $\models$ inductively as follows (clauses for Boolean connectives are omitted):*

$$\begin{array}{lll} (\mathcal{M}, \mathbf{c}) \models F_a p & \textit{iff} & p \in \mathbf{c}(\mathsf{root})_a, \\ (\mathcal{M}, \mathbf{c}) \models K_a \phi & \textit{iff} & \forall \mathbf{d} \ \textit{s.t.}\ \mathbf{c} \sim_a \mathbf{d},\ (\mathcal{M}, \mathbf{d}) \models \phi. \end{array}$$

So formula $F_a p$ is true (in a pointed gossip model) whenever secret p belongs to the set of secrets agent a is familiar with in the situation generated by the designated call sequence $\mathbf{c}$ applied to the initial situation root. The knowledge operator is interpreted as customary in epistemic logic using the equivalence relations $\sim_a$.

[2] Notice that the definition requires a designated initial situation, which we assume to be root.

2.2.5 Computations

Assume a gossip protocol P that is a parallel composition of the component programs $*[[]_{j=1}^{m_a}\ \psi_j^a \to \mathsf{c}_j^a]$, one for each agent $a \in \mathsf{A}$.

Given the gossip model $\mathcal{M} = (\mathbf{C}^{<\omega}, \{\sim_a\}_{a\in\mathsf{A}})$ we define the ***computation tree*** $\mathbf{C}^P \subseteq \mathbf{C}^{<\omega}$ of P as the smallest set of sequences satisfying the following inductive conditions:

[Base] $\epsilon \in \mathbf{C}^P$;

[Step] If $\mathbf{c} \in \mathbf{C}^P$ and $(\mathcal{M}, \mathbf{c}) \models \psi_j^a$ then $\mathbf{c}.\mathsf{c}_j^a \in \mathbf{C}^P$. In this case we say that a ***transition*** has taken place between $\mathbf{c}$ and $\mathbf{c}.\mathsf{c}_j^a$, in symbols, $\mathbf{c} \to \mathbf{c}.\mathsf{c}_j^a$.

So $\mathbf{C}^P$ is a (possibly infinite) set of finite call sequences that is iteratively obtained by performing a 'legal' call (according to protocol P) from a 'legal' (according to protocol P) call sequence.

A ***path*** in the computation tree of P is a (possibly infinite) sequence of elements of $\mathbf{C}^P$, denoted by $\xi = (\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_n, \ldots)$, where $\mathbf{c}_0 = \epsilon$ and each $\mathbf{c}_{i+1} = \mathbf{c}_i.\mathsf{c}$ for some call c and $i \geq 0$. A ***computation*** of P is a maximal rooted path in the computation tree of P.[3]

The above definition implies that a call sequence $\mathbf{c}$ is a leaf of the computation tree if and only if

$$(\mathcal{M}, \mathbf{c}) \models \bigwedge_{a\in\mathsf{A}} \bigwedge_{j=1}^{m_a} \neg\psi_j^a.$$

We call the formula

$$\bigwedge_{a\in\mathsf{A}} \bigwedge_{j=1}^{m_a} \neg\psi_j^a$$

the ***exit condition*** of the gossip protocol P.

Obviously computation trees can be infinite, though they are always finitely branching. Further, note that this semantics for gossip protocols abstracts away from some implementation details of the calls. More specifically, we assume that the caller always succeeds in his call and does not require to synchronize with the called agent. In reality, the called agent might be busy, being engaged in another call. To take care of this one could modify each call by replacing it by a 'call protocol' that implements the actual call using some lower level primitives. We do not elaborate further on this topic.

Let us fix some more terminology. For $\mathbf{c} \in \mathbf{C}^P$, an agent a is ***enabled*** in $\mathbf{c}$ if $(\mathcal{M}, \mathbf{c}) \models \bigvee_{j=1}^{m_a}\ \psi_j^a$ and is ***disabled*** otherwise. So an agent is enabled if it can perform a call. An agent a is ***selected*** in $\mathbf{c}$ if it is the caller in the call that for some $\mathbf{c}'$ determines the transition $\mathbf{c} \to \mathbf{c}'$ in ξ. Finally, a computation ξ is called a ***fair computation*** if it is finite or each agent that is enabled in infinitely many sequences in ξ is selected in infinitely many sequences in ξ.

We note in passing that various alternative definitions of fairness are possible; we just focus on one of them. An interested reader may consult [2], where several fairness definitions (for instance one focusing on actions and not on agents) for distributed programs were considered and compared.

We conclude this section by observing the following. Our definition of computation tree for protocol P presupposes that guards ψ_j^a are interpreted over the gossip model $\mathcal{M} = (\mathbf{C}^{<\omega}, \{\sim_a\}_{a\in\mathsf{A}})$. This means that when evaluating guards, agents consider as possible call sequences that cannot be generated by P. In other words, agents do not know the protocol. To model common knowledge of the considered protocol in the gossip model one should take as the domain of the gossip model $\mathcal{M}$ the underlying computation tree. However, the computation tree is defined by means of the underlying gossip model. To handle such a circularity an appropriate fixpoint definition is needed. We leave this topic for future work.

2.3 Correctness

We are interested in proving the correctness of gossip protocols. Assume a gossip protocol P that is a parallel composition of the component programs $*[[]_{j=1}^{m_a}\ \psi_j^a \to \mathsf{c}_j^a]$.

We say that P is ***partially correct***, in short ***correct***, if in all situations sequences $\mathbf{c}$ that are leaves of the computation tree of P, for each agent a

$$(\mathcal{M}, \mathbf{c}) \models \bigwedge_{b\in\mathsf{A}} F_a B,$$

i.e., if for all situations sequences $\mathbf{c}$ that are leaves of the computation tree of P, each agent is an expert in the gossip situation $\mathbf{c}(\mathsf{root})$.

We say furthermore that P ***terminates*** if all its computations are finite and that P ***fairly terminates*** if all its fair computations are finite.

In the next section we provide examples showing that partial correctness and termination of the considered protocols can depend on the assumed mode of communication and on the number of agents. In what follows we study various gossip protocols and their correctness. We begin with the following obvious observation.

FACT 2.7. *For each protocol P the following implications ($\Rightarrow$) hold, where $T_P(x)$ stands for its termination and $FT_P(x)$ for its fair termination in a communication mode x:*

$$T_P(x) \Rightarrow FT_P(x).$$

Protocol R3 given in Section 4 shows that none of these implications can be reversed. Moreover, it is not the case either that for each protocol P:

$$T_P(\triangleright) \Rightarrow T_P(\text{push-pull}),$$
$$T_P(\triangleleft) \Rightarrow T_P(\text{push-pull}).$$

EXAMPLE 2.8. *Let $\mathsf{A} = \{a, b, c\}$ and define the following expression:*

$$\mathcal{A} \subset \mathcal{C} \quad := \quad \bigwedge_{I\in\{A,B,C\}} (F_a I \to F_c I) \wedge \bigvee_{I\in\{A,B,C\}} (F_c I \wedge \neg F_a I)$$

Expression $\mathcal{B} \subset \mathcal{C}$ can be defined analogously. Note that we denote by I the secret of agent i. Intuitively, $\mathcal{A} \subset \mathcal{C}$ means that agent c is familiar with all the secrets that agent a is familiar with, but not vice versa. So c is familiar with a superset of the secrets a is aware of. Further, let Exp_j stand for $\bigwedge_{I\in\{A,B,C\}} F_j I$.

Consider now the following component programs:

- *for agent a:* $*[K_a(\neg(\mathcal{A} \subset \mathcal{C}) \wedge \neg Exp_a) \to a \triangleright c]$,

[3] Note that while the sequences that are elements of the computation tree of a protocol are always finite (although possibly infinite in number), computations can be infinite sequences (of finite call sequences).

- *for agent b:* $*[K_b(\neg(\mathcal{B} \subset \mathcal{C}) \wedge \neg Exp_b) \rightarrow b \triangleright c]$,
- *for agent c:* $*[[]_{i\in\{a,b\}} K_c Exp_c \wedge \neg K_c Exp_i \rightarrow c \triangleright i]$.

First note that in our logic, $K_i(\phi_1 \wedge \phi_2)$ *is equivalent to* $K_i\phi_1 \wedge K_i\phi_2$.

This protocol is correct. Indeed, initially, it is not the case that c knows to be an expert, hence the guard of c is false. Likewise, the guards of a and b are true; a for instance knows that c is not familiar with more secrets than a, and that a is not familiar with all secrets. So initially both a or b are enabled. If the first call is granted to a, this agent will call c, yielding the situation A.B.AC. Note that now, the guard of a is false (from a's perspective, c may now well be familiar with all secrets), the guard of b is true, and the guard of c is still false. So now only b is enabled, which yields the situation A.B.ABC. At this stage, only agent c is enabled and after he calls both a and b all guards become false.

Moreover, this protocol terminates. Indeed, the only computations are the ones in which first the calls $a \triangleright c$ *and* $b \triangleright c$ *take place, in any order, followed by the calls* $c \triangleright a$ *and* $c \triangleright b$*, also performed in any order. However, if we use the push-pull mode instead of push, the call ac can be indefinitely repeated, so the protocol does not terminate.* □

3. TWO SYMMETRIC PROTOCOLS

In this section we consider protocols for the case when the agents form a complete graph. We study two protocols. We present them first for the communication mode push-pull. (Partial) correctness of the considered protocols does not depend on the assumed mode of communication.

Learn new secrets protocol (LNS).

Consider the following program for agent i:

$$*[[]_{j\in\mathsf{A}} \neg F_i J \rightarrow (i,j)].$$

Informally, agent i calls agent j if i is not familiar with j's secret. Note that the guards of this protocol do not use the epistemic operator K_i, but they are equivalent to the ones that do, as $\neg F_i J$ is equivalent to $K_i \neg F_i J$.

This protocol was introduced in [3] and studied with respect to the push-pull mode, assuming asynchronous communication. As noted there this protocol is clearly correct. Also, it always terminates since after each call (i,j) the size of $\{(i,j) \in \mathsf{A} \times \mathsf{A} \mid \neg F_i J\}$ decreases. The same argument shows termination if the communication mode is pull.

However, if the communication mode is push, the protocol may fail to terminate, even fairly. To see it fix an agent a and consider a sequence of calls in which each agent calls a. At the end of this sequence a becomes an expert but nobody is familiar with his secret. So any extension of this sequence is an infinite computation.

Let us consider now the possible call sequences generated by the computations of this protocol. Assume that there are $n \geq 4$ agents. By the result mentioned in the introduction in each terminating computation at least $2n-4$ calls are made.

The LNS protocol can generate such shortest sequences (among others). Indeed, let $\mathsf{A} = \{a, b, c, d, i_1, \dots, i_{n-4}\}$ be the set of agents. Then the following sequence of $2n-4$ calls

$$\begin{gathered}(a, i_1), (a, i_2), \dots, (a, i_{n-4}), \\ (a, b), (c, d), (a, c), (b, d), \\ (i_1, b), (i_2, b), \dots, (i_{n-4}, b)\end{gathered} \tag{1}$$

corresponds to a terminating computation.

The guards used in this protocol entail that after a call (i,j) neither the call (j,i) nor another call (i,j) can take place, that is between each pair of agents at most one call can take place. Consequently, the longest possible sequence contains at most $\frac{n(n-1)}{2}$ calls. Such a worst case can be generated by means of the following sequence of calls:

$$[2],\ [3],\ [4], \dots, [n],$$

where for a natural number k, $[k]$ stands for the sequence $(1,k)$, $(2,k)$, $\dots$, $(k-1,k)$.[4]

Hear my secret protocol (HMS).

Next, we consider a protocol with the following program for agent i:

$$*[[]_{j\in\mathsf{A}} \neg K_i F_j I \rightarrow (i,j)].$$

Informally, agent i calls agent j if he (agent i) does not know whether j is familiar with his secret. To prove correctness of this protocol it suffices to note that its exit condition

$$\bigwedge_{i,j\in\mathsf{A}} K_i F_j I$$

implies $\bigwedge_{i,j\in\mathsf{A}} F_j I$. To prove termination it suffices to note that after each call (i,j) the size of the set $\{(i,j) \mid \neg K_i F_j I\}$ decreases.

If the communication mode is push, then the termination argument remains valid, since after the call $i \triangleright j$ agent j still learns all the secrets agent i is familiar with.

However, if the communication mode is pull, then the protocol may fail to terminate, even fairly. To see it fix an agent j and consider the calls $i \triangleleft j$, where i ranges over $\mathsf{A} \setminus \{j\}$, arbitrarily ordered. Denote this sequence by $\mathbf{c}$. Consider now an infinite sequence of calls resulting from repeating $\mathbf{c}$ indefinitely. It is straightforward to check that such a sequence corresponds to a possible computation. Indeed, in this sequence agent j never calls and hence never learns any new secret. So for each $i \neq j$ the formula $\neg K_i F_j I$ remains true and hence each agent $i \neq j$ remains enabled. Moreover, after the calls from $\mathbf{c}$ took place agent j is not anymore enabled. Hence the resulting infinite computation is fair.

When there are $n \geq 4$ agents, the extreme cases in terms of the lengths of possible call sequences are the same as in the case of the LNS protocol. Indeed, let $\mathsf{A} = \{a, b, c, d, i_1, \dots, i_{n-4}\}$ be the set of agents. Then the sequence of (1) corresponds to a terminating computation. Further, this protocol can generate computations in which $\frac{n(n-1)}{2}$ calls are made. The argument is the same as for the LNS protocol.

4. PROTOCOLS OVER DIRECTED RINGS

In this section we consider the case when the agents are arranged in a directed ring, where $n \geq 3$. For convenience we take the set of agents to be $\{1, 2, \dots, n\}$. For $i \in \{1, \dots, n\}$, let $i \oplus 1$ and $i \ominus 1$ denote respectively the successor and predecessor of agent i. That is, for $i \in \{1, \dots, n-1\}$, $i \oplus 1 = i + 1$, $n \oplus 1 = 1$, for $i \in \{2, \dots, n\}$, $i \ominus 1 = i - 1$, and $1 \ominus 1 = n$. For $k > 1$ we define $i \oplus k$ and $i \ominus k$ by induction in the expected way. Again, when reasoning about the protocols we denote the secret of agent $i \in \{1, \dots, n\}$ by

[4] Other longest sequences are obviously possible, for instance: $12, 13, ..., 1n, 23, 24, ..., 2n, 34, 35, .., 3n, ..., (n-1)n$.

I. We consider four different protocols and study them with respect to their correctness and (fair) termination.

In this set up, a call sequence over a directed ring is a (possibly infinite) sequence of calls, all being of the same communication mode, and all involving an agent i and $i \oplus 1$. As before, we use $\mathbf{c}$ to denote such a call sequence and $\mathbf{C}_{DR}$ to denote the set of all call sequences over a directed ring. In this section, unless stated otherwise, by a call sequence we mean a sequence over a directed ring. The set of all such finite call sequences is denoted $\mathbf{C}_{DR}^{<\omega}$. A gossip model for a directed ring is a tuple $\mathcal{M}_{DR} = (\mathbf{C}_{DR}^{<\omega}, \{\sim_a\}_{a \in \mathsf{A}})$, where each $\sim_a \subseteq \mathbf{C}_{DR}^{<\omega} \times \mathbf{C}_{DR}^{<\omega}$ is as in Definition 2.4. The truth definition is as before, and the notion of a ***computation tree for directed rings*** $\mathbf{C}_{DR}^{P} \subseteq \mathbf{C}_{DR}^{<\omega}$ of a ring protocol P is analogous to the notion defined before. Note that by restricting the domain in $\mathcal{M}_{DR}$ to $\mathbf{C}_{DR}^{<\omega}$, the ring network—and hence who is the successor of whom—becomes common knowledge.

When presenting the protocols we use the fact that $F_i J$ is equivalent to $K_i F_i J$.

Ring protocol R1.

Consider first a gossip protocol with the following program for i:

$$*[\bigvee_{j=1}^{n} (F_i J \wedge K_i \neg F_{i\oplus 1} J) \rightarrow i \diamond i \oplus 1],$$

where $\diamond$ denotes the mode of communication, so $\rhd$, $\lhd$ or push-pull.

Informally, agent i calls his successor, agent $i \oplus 1$, if i is familiar with some secret and he knows that his successor is not familiar with it.

Proposition 4.1. *Let $\diamond = \rhd$. Protocol R1 terminates and is correct.*

Termination and correctness do not both hold for the other communication modes. Consider first the pull communication mode, i.e., $\diamond = \lhd$. Then the protocol does not always terminate. Indeed, each call $i \lhd i \oplus 1$ can be repeated. Next, consider the push-pull communication mode. We show that then the protocol is not correct. Indeed, take

$$\mathbf{c} = (1,2),\ (2,3), \ldots, (n-1,n).$$

We claim that after the sequence of calls $\mathbf{c}$ the exit condition of the protocol is true. To this end we consider each agent in turn.

After $\mathbf{c}$ each agent i, where $i \neq n$ is familiar the secrets of the agents $1, 2, \ldots, i+1$. Moreover, because of the call $(i, i+1)$ agent i knows that agent $i+1$ is familiar with these secrets. So the exit condition of agent i is true.

To deal with agent n note that $\mathbf{c} \sim_n \mathbf{c}.(n-2, n-1).(n-3, n-2). \ldots .(2,3).(1,2)$. After the latter call sequence agent 1 becomes an expert. So after $\mathbf{c}$ agent n cannot know that agent 1 is not familiar with some secret. Consequently, after $\mathbf{c}$ the exit condition of agent n is true, as well. However, after $\mathbf{c}$ agent 1 is not an expert, so the protocol is indeed not correct.

In what follows we initially present the protocols assuming the push-pull mode of communication.

Ring protocol R2.

Consider now a gossip protocol with the following program for agent i:

$$*[\neg K_i F_{i\oplus 1} I \ominus 1 \rightarrow (i, i \oplus 1)],$$

where (recall) $I \ominus 1$ denotes the secret of agent $i \ominus 1$. Informally, agent i calls his successor, agent $i \oplus 1$, if i does not know that his successor is familiar with the secret of i's predecessor, i.e., agent $i \ominus 1$.

Proposition 4.2. *If $|\mathsf{A}| \in \{3, 4\}$ then protocol R2 is correct.*

However, this protocol is not correct for five or more agents. To see it consider the sequence of calls

$$(1,2),\ (2,3),\ \ldots, (n-1,n),\ (n,1),\ (1,2)$$

where $n \geq 5$. After it the exit condition of the protocol is true. However, agent 3 is not familiar with the secret of agent 5.

Note that the same argument shows that the protocol in which we use $\neg K_i F_{i\oplus 1} I \vee \neg K_i F_{i\oplus 1} I \ominus 1$ instead of $\neg K_i F_{i\oplus 1} I \ominus 1$ is incorrect, as well.

Moreover, this protocol does not always terminate. Indeed, one possible computation consists of an agent i repeatedly calling his successor $i \oplus 1$.

Ring protocol R3.

Next, consider the following modification of protocol R2 in which we use the following program for agent i:

$$*[(\neg \bigwedge_{j=1}^{n} F_i J) \vee \neg K_i F_{i\oplus 1} I \ominus 1 \rightarrow (i, i \oplus 1)].$$

Informally, agent i calls his successor, agent $i \oplus 1$, if i is not familiar with all the secrets or i does not know that his successor is familiar with the secret of his predecessor, agent $i \ominus 1$.

This gossip protocol is obviously correct thanks to the fact that $\bigwedge_{i=1}^{n} \bigwedge_{j=1}^{n} F_i J$ is part of the exit condition. However, it does not always terminate for the same reason as the previous one.

On the other hand, the following holds.

Proposition 4.3. *Protocol R3 fairly terminates.*

The same conclusions concerning non termination and fair termination can be drawn for the push and the pull modes of communication. Indeed, for push it suffices to consider the sequence of calls $i \rhd i \oplus 1,\ i \oplus 1 \rhd i \oplus 2, \ldots, i \ominus 1 \rhd i$ after which agent $i \ominus 1$ becomes disabled, and for pull the sequence of calls $i \lhd i \oplus 1,\ i \ominus 1 \lhd i, \ldots, i \oplus 2 \lhd i \oplus 3$ after which agent $i \oplus 2$ becomes disabled.

Ring protocol R4.

Finally, we consider a protocol that is both correct and terminates for the push-pull mode. Consider the following program for i:

$$*[\bigvee_{j=1}^{n} (F_i J \wedge \neg K_i F_{i\oplus 1} J) \rightarrow (i, i \oplus 1)].$$

Informally, agent i calls his successor, agent $i \oplus 1$, if i is familiar with some secret and he does not know whether his successor is familiar with it. Note the similarity with protocol R1.

Protocol	T	FT	T for $\rhd$	FT for $\rhd$	T for $\lhd$	FT for $\lhd$
LNS	yes	yes	no	no	yes	yes
HMS	yes	yes	yes	yes	no	no
R3	no	yes	no	yes	no	yes
R4	yes	yes	yes	yes	no	yes

Table 1: Summary of termination results.

Proposition 4.4. *Protocol R4 terminates and is correct.*

If the communication mode is push, then the termination argument remains valid, since after the call $i \rhd i \oplus 1$ agent $i \oplus 1$ still learns all the secrets that agent i is familiar with and hence the above set $\{(i,j) \mid \neg K_i F_{i\oplus 1} J\}$ decreases.

If the communication mode is pull, then the protocol may fail to terminate, because after the first call $i \lhd i \oplus 1$ agent $i \oplus 1$ does not learn the secret of agent i and consequently the call can be repeated. However, the situation changes when fairness is assumed.

Proposition 4.5. *For the pull communication mode protocol R4 fairly terminates.*

Table 1 summarizes the termination properties of the protocols considered in the paper.

5. CONCLUSIONS

The aim of this paper was to introduce distributed gossip protocols, to set up a formal framework to reason about them, and to illustrate it by means of an analysis of selected protocols.

Our results open up several avenues for further research. First, our correctness arguments were given in plain English with occasional references to epistemic tautologies, such as $K_i\phi \to \phi$, but it should be possible to formalize them in a customized epistemic logic. Such a logic should have a protocol independent component that would consist of the customary S5 axioms and a protocol dependent component that would provide axioms that depend on the mode of communication and the protocol in question. An example of such an axiom is the formula $K_iF_{i\oplus 1}I \ominus 1 \to F_iI \oplus 1$ that we used when reasoning about protocol R2. To prove the validity of the latter axioms one would need to develop a proof system that allows us to compute the effect of the calls, much like the computation of the strongest postconditions in Hoare logics. Once such a logic is provided the next step will be to study formally its properties, including decidability. Then we could clarify whether the provided correctness proofs could be carried out automatically.

Second, generalizing further the ideas we introduced by considering directed rings, gossip protocols could be studied in interface with network theory (see [13] for a textbook presentation). Calls can be assumed to be constrained by a network, much like in the literature on 'centralized' gossip (cf. [10]) or even have probabilistic results (i.e., secrets are passed with given probabilities). More complex properties of gossip protocols could then be studied involving higher-order knowledge or forms of group knowledge among neighbors (e.g., "it is common knowledge among a and her neighbors that they are all experts"), or their stochastic behavior (e.g., "at some point in the future all agents are experts with probability p").

Third, it will be interesting to analyze the protocols for the types of calls considered in [3]. They presuppose some form of knowledge that a call took place (for instance that given a call between a and b each agent $c \neq a, b$ noted the call but did not learn its content). Another option is to consider multicasting (calling several agents at the same time).

Finally, many assumptions of the current setup could be lifted. Different initial and final situations could be considered, for instance common knowledge of protocols could be assumed, or common knowledge of the familiarity of all agents with all the secrets upon termination could be required. Finally, to make the protocols more efficient passing of tokens could be allowed instead of just the transmission of secrets by means of calls.

Acknowledgments

We would like to thank Hans van Ditmarsch and the referees for helpful comments and Rahim Ramezanian for useful comments about Example 2.8. This work resulted from a research visit by Krzysztof Apt to Davide Grossi and Wiebe van der Hoek, sponsored by the 2014 Visiting Fellowship Scheme of the Department of Computer Science of the University of Liverpool. The first author is also a Visiting Professor at the University of Warsaw. He was partially supported by the NCN grant nr 2014/13/B/ST6/01807.

6. REFERENCES

[1] K. R. Apt, F. R. de Boer, and E. R. Olderog. *Verification of Sequential and Concurrent Programs.* Springer, 2009.

[2] K. R. Apt, N. Francez, and S. Katz. Appraising fairness in distributed languages. *Distributed Computing*, 2(4):226–241, 1988.

[3] M. Attamah, H. van Ditmarsch, D. Grossi, and W. Van der Hoek. Knowledge and gossip. In *Proceedings of ECAI'14*, pages 21–26. IOS Press, 2014.

[4] B. Baker and R. Shostak. Gossips and telephones. *Discrete Mathematics*, 2:197–193, 1972.

[5] R. Bumby. A problem with telephones. *SIAM Journal of Algorithms and Discrete Methods*, 2:13–18, 1981.

[6] E. W. Dijkstra. Guarded commands, nondeterminacy and formal derivation of programs. *Communications of the ACM*, 18:453–457, 1975.

[7] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. Knowledge-based programs. *Distributed Computing*, 10:199–225, 1997.

[8] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about knowledge.* The MIT Press, Cambridge, 1995.

[9] A. Hajnal, E. C. Milner, and E. Szemeredi. A cure for the telephone disease. *Canadian Mathematical Bulletin*, 15:447–450, 1972.

[10] S. M. Hedetniemi, S. T. Hedetniemi, and A. L. Liestman. A survey of gossiping and broadcasting in communication networks. *Networks*, 18(4):319–349, 1988.

[11] C. A. R. Hoare. Communicating sequential processes. *Communications of the ACM*, 21:666–677, 1978.

[12] INMOS Limited. *Occam Programming Manual.* Prentice-Hall International, 1984.

[13] M. O. Jackson. *Social and Economic Networks.* Princeton University Press, 2008.

[14] R. Kurki-Suonio. Towards programming with knowledge expressions. In *Proceedings of POPL'86*, pages 140–149, 1986.

[15] J. Meyer and W. van der Hoek. *Epistemic Logic for AI and Computer Science*, volume 41 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1995.

[16] R. Parikh and R. Ramanujam. Distributed processing and the logic of knowledge. In *Logic of Programs*, LNCS 193, pages 256–268. Springer, 1985. Similar to *JoLLI* 12: 453–467, 2003.

[17] A. Seress. Quick gossiping without duplicate transmissions. *Graphs and Combinatorics*, 2:363–383, 1986.

[18] R. Tijdeman. On a telephone problem. *Nieuw Archief voor Wiskunde*, 3(XIX):188–192, 1971.

APPENDIX

PROOF OF PROPOSITION 4.1.
Termination Given a call sequence c define the set

$$Inf(\mathsf{c}) := \{(i,j) \mid i,j \in \{1,\ldots,n\} \text{ and } (\mathcal{M}_{DR},\mathsf{c}) \models F_iJ\}.$$

After each enabled call $i \triangleright i \oplus 1$ in c, the set $Inf(\mathsf{c})$ increases, which ensures termination since each set $Inf(\cdot)$ has at most n^2 elements.
Correctness Consider a leaf of the computation tree. Then the exit condition

$$\bigwedge_{i=1}^{n}\bigwedge_{j=1}^{n}(\neg F_iJ \vee \neg K_i\neg F_{i\oplus 1}J)$$

is true. We proceed by induction to show that then each F_iJ is true, where $i,j \in \{1,\ldots,n\}$, and where the pairs (i,j) are ordered as follows:

$$\begin{aligned}&(1,1),(2,1),\ldots,(n,1),\\&(2,2),(3,2),\ldots,(1,2),\\&\ldots,\\&(n,n),(1,n),\ldots,(n-1,n).\end{aligned}$$

So the ith row lists the pairs (j,i) with $j \in \{1,\ldots,n\}$ ranging clockwise, starting at i.

Take a pair (i,j). If $i = j$, then F_iJ is true by assumption. If $i \neq j$, then consider the pair that precedes it in the above ordering. It is then of the form (i_1,j), where $i = i_1 \oplus 1$. By the induction hypothesis $F_{i_1}J$ is true, so by the exit condition $\neg K_{i_1}\neg F_iJ$ is true.

Suppose now towards a contradiction that $\neg F_{i_1\oplus 1}J$ is true. Then $i_1 \oplus 1 \neq j$. Hence by virtue of the considered communication mode and Definition 2.4 it follows that agent i_i knows that $\neg F_{i_1\oplus 1}J$ is true since the only way for $i_1 \oplus 1$ to become familiar with J is by means of a call from i_1. So $K_{i_1}\neg F_iJ$ is true. This yields a contradiction. Hence F_iJ is true.

So we showed, as desired, that $\bigwedge_{i=1}^{n}\bigwedge_{j=1}^{n} F_iJ$ is true in the considered leaf. □

PROOF OF PROPOSITION 4.2. To start with, $\bigwedge_{i=1}^{n} F_iI$ is true in every node of the computation tree. Suppose the exit condition $\bigwedge_{i=1}^{n} K_iF_{i\oplus 1}I \ominus 1$ is true at a node of the computation tree (in short, true). It implies that $\bigwedge_{i=1}^{n} F_{i\oplus 1}I \ominus 1$ is true. Fix $i \in \{1,\ldots,n\}$. By the above $F_iI \ominus 2$ is true. Further, the implication $K_iF_{i\oplus 1}I \ominus 1 \rightarrow F_iI \ominus 1$ is true in every node of the computation tree (remember, the agents are positioned on a directed ring). If $n = 3$, this proves that $\bigwedge_{j=1}^{n} F_iJ$ is true.
If $n = 4$, we note that $K_iF_{i\oplus 1}I \ominus 1$ implies that agent $i \oplus 1$ learned $I \ominus 1$ through a call of agent i and hence the implication $K_iF_{i\oplus 1}I \ominus 1 \rightarrow F_iI \oplus 1$ is true in every node of the computation tree, as well (remember that the mode is push-pull). We conclude that $\bigwedge_{j=1}^{n} F_iJ$ is true. □

PROOF OF PROPOSITION 4.3. First, note that the following three statements are equivalent for each node c of an arbitrary computation ξ and each agent i:

- i is disabled at c,
- $(\mathcal{M}_{DR},\mathsf{c}) \models (\bigwedge_{j=1}^{n} F_iJ) \wedge K_iF_{i\oplus 1}I \ominus 1$,
- a sequence of calls $(i\oplus 2, i\oplus 3)$, $(i\oplus 3, i\oplus 4),\ldots,(i, i\oplus 1)$ (possibly interspersed with other calls) has taken place in ξ before c.

Suppose now towards a contradiction that an infinite fair computation ξ exists. We proceed by case distinction.
Case 1 Some agent becomes disabled in ξ.
We claim that if an agent i becomes disabled in ξ, then also agent $i \oplus 1$ becomes disabled in ξ. Indeed, otherwise by fairness at some point in ξ after which i becomes disabled, agent $i\oplus 1$ calls his successor, $i\oplus 2$, and by the above sequence of equivalences in turn becomes disabled.

We conclude by induction that at some point in ξ all agents become disabled and hence ξ terminates, which yields a contradiction.
Case 2 No agent becomes disabled in ξ.
By fairness each agent calls in ξ infinitely often his successor. So for every agent i there exists in ξ the sequence of calls $(i\oplus 2, i\oplus 3)$, $(i\oplus 3, i\oplus 4),\ldots,(i, i\oplus 1)$ (possibly interspersed with other calls). By the above sequence of equivalences after this sequence of calls agent i becomes disabled, which yields a contradiction. □

PROOF OF PROPOSITION 4.4.
Termination It suffices to note that after each call $(i, i\oplus 1)$ the size of the set

$$\{(i,j) \in \mathsf{A} \times \mathsf{A} \mid \neg K_iF_{i\oplus 1}J\}$$

decreases.
Correctness Consider a leaf of the computation tree. Then the exit condition

$$\bigwedge_{i=1}^{n}\bigwedge_{j=1}^{n}(\neg F_iJ \vee K_iF_{i\oplus 1}J)$$

is true. As in the case of protocol R1 we prove that it implies each F_iJ is true by induction on the pairs (i,j), where $i,j \in \{1,\ldots,n\}$, ordered as follows:

$$\begin{aligned}&(1,1),(2,1),\ldots,(n,1),\\&(2,2),(3,2),\ldots,(1,2),\\&\ldots,\\&(n,n),(1,n),\ldots,(n-1,n).\end{aligned}$$

Take a pair (i,j). If $i = j$, then F_iJ is true by assumption. If $i \neq j$, then consider the pair that precedes it in the above ordering, so (i_1,j), where $i = i_1 \oplus 1$. By the induction hypothesis $F_{i_1}J$ is true, so by the exit condition $K_{i_1}F_iJ$ is true and hence F_iJ is true. □

Proof of Proposition 4.5. Consider the following sequence of statements:

(i) i is disabled at $\mathbf{c}$,

(ii) $(\mathcal{M}_{DR}, \mathbf{c}) \models \bigwedge_{j=1}^{n}(F_i J \to K_i F_{i\oplus 1} J)$,

(iii) $(\mathcal{M}_{DR}, \mathbf{c}) \models K_i F_{i\oplus 1}$,

(iv) a sequence of calls $i \ominus 1 \lhd i, i \ominus 2 \lhd i \ominus 1, \ldots, i \lhd i \oplus 1$ (possibly interspersed with other calls) has taken place in ξ before $\mathbf{c}$.

It is easy to verify that these statements are logically related in the following way:

$$\text{(i)} \Leftrightarrow \text{(ii)} \Rightarrow \text{(iii)} \Rightarrow \text{(iv)} \Rightarrow \text{(ii)}$$

for each node $\mathbf{c}$ of an arbitrary computation ξ and each agent i. They are therefore equivalent. Suppose now towards a contradiction that an infinite fair computation ξ exists. As in the proof of Proposition 4.3 we proceed by case distinction.

$\boxed{\text{Case 1}}$ Some agent becomes disabled in ξ.
We claim that if an agent i becomes disabled in ξ, then also $i \ominus 1$ becomes disabled in ξ. Indeed, otherwise by fairness at some point in ξ after which j becomes disabled, agent $i \ominus 1$ calls his successor, i, and by the above sequence of equivalences in turn becomes disabled.

We conclude by induction that at some point in ξ all agents become disabled and hence ξ terminates, which yields a contradiction.

$\boxed{\text{Case 2}}$ No agent becomes disabled in ξ.
By fairness each agent calls in ξ infinitely often his successor. So for every agent i there exists in ξ a sequence of calls $i \ominus 1 \lhd i, i \ominus 2 \lhd i \ominus 1, \ldots, i \lhd i \oplus 1$ (possibly interspersed with other calls). By the above sequence of equivalences, after this sequence of calls agent i becomes disabled, which yields a contradiction. $\square$

Coordination Games on Directed Graphs

Krzysztof R. Apt
Centrum Wiskunde &
Informatica
Amsterdam, The Netherlands
k.r.apt@cwi.nl

Sunil Simon
Department of Computer
Science and Engineering
IIT Kanpur, Kanpur, India
simon@cse.iitk.ac.in

Dominik Wojtczak
University of Liverpool
Liverpool, U.K.
d.wojtczak@liv.ac.uk

ABSTRACT

We study natural strategic games on directed graphs, which capture the idea of coordination in the absence of globally common strategies. We show that these games do not need to have a pure Nash equilibrium and that the problem of determining their existence is NP-complete. The same holds for strong equilibria. We also exhibit some classes of games for which strong equilibria exist and prove that a strong equilibrium can then be found in linear time.

1. INTRODUCTION

In this paper we study a simple and natural class of strategic games. Assume a finite directed graph. Suppose that each node selects a colour from a private set of colours available for it. The payoff to a node is the number of (in)neighbours who chose the same colour.

These games are typical examples of coordination games. Recall that the idea behind *coordination* in strategic games is that players are rewarded for choosing common strategies. The games we study here are specific coordination games in the absence of globally common strategies.

Recently, we studied in [2], and more fully in [3], a very similar class of games in which the graphs were assumed to be undirected. However, the transition from undirected to directed graphs drastically changes the status of the games. For instance, for the case of directed graphs Nash equilibria do not need to exist, while they always exist when the graph is undirected. Consequently, in [2] and [3] we focused on the problem of existence of strong equilibria. We also argued there that such games are of relevance for the cluster analysis, the task of which is to partition in a meaningful way the nodes of a graph. The same applies here. Indeed, once the strategies are possible cluster names, a Nash equilibrium naturally corresponds to a ‘satisfactory’ clustering of the underlying graph.

The above two classes of games are also similar in that both are special cases of a number of well-studied types of games. One of them are ***polymatrix games*** introduced in [10]. In these games the payoff for each player is the sum of the payoffs from the individual two player games he plays with each other player separately. Another are ***graphical games*** introduced in [11]. In these games the payoff of each player depends only on the strategies of its neighbours in a given in advance graph structure over the set of players.

In addition both classes of games satisfy the ***positive population monotonicity*** (PPM) property introduced in [12] that states that the payoff of each player weakly increases if another player switches to his strategy. Coordination games on graphs are examples of games on networks, a vast research area recently surveyed in [9]. Other related references can be found in [3].

1.1 Plan of the paper and the results

In the next section we introduce preliminary definitions, following [3]. We define the coordination games on directed graphs in Section 3. In Section 4 we exhibit a number of cases when a strong equilibrium exists. Next, in Section 5 we study complexity of the problem of existence of Nash and strong equilibria and the problem of determining the complexity of finding a strong equilibrium in a natural case when it is known to exist. Finally, in Section 6 we discuss future directions.

The main results are as follows. If the underlying graph is a DAG, is complete or is such that every strongly connected component (SCC) is a simple cycle, then strong equilibria always exist and they can always be reached from any initial joint strategy by a sequence of coalitional improvement steps. The same is the case when only two colours are used.

In general Nash equilibria do not need to exist and the problem of determining their existence is NP-complete. The same is the case for strong equilibria. We also show that when every SCC is a simple cycle, then strong equilibrium can always be found in linear time.

2. PRELIMINARIES

A ***strategic game*** $\mathcal{G} = (S_1, \ldots, S_n, p_1, \ldots, p_n)$ with $n > 1$ players, consists of a non-empty set S_i of ***strategies*** and a ***payoff function*** $p_i : S_1 \times \cdots \times S_n \to \mathbb{R}$, for each player i.

We denote $S_1 \times \cdots \times S_n$ by S, call each element $s \in S$ a ***joint strategy*** and abbreviate the sequence $(s_j)_{j \neq i}$ to s_{-i}. Occasionally we write (s_i, s_{-i}) instead of s. We call a strategy s_i of player i a ***best response*** to a joint strategy s_{-i} of his opponents if for all $s'_i \in S_i$, $p_i(s_i, s_{-i}) \geq p_i(s'_i, s_{-i})$.

Fix a strategic game $\mathcal{G}$. We say that $\mathcal{G}$ satisfies the ***positive population monotonicity (PPM)*** if for all joint strategies s and players i, j, $p_i(s) \leq p_i(s_i, s_{-j})$. (Note that (s_i, s_{-j}) refers to the joint strategy in which player j chooses s_i.) So if more players (here just player j) choose player i's strategy and the remaining players do not change their strategies, then i's payoff weakly increases.

We call a non-empty subset $K := \{k_1, \ldots, k_m\}$ of the set of players $N := \{1, \ldots, n\}$ a ***coalition***. Given a joint strategy s we abbreviate the sequence $(s_{k_1}, \ldots, s_{k_m})$ of strategies to s_K and $S_{k_1} \times \cdots \times S_{k_m}$ to S_K. We also write (s_K, s_{-K}) instead of s. If there is a strategy x such that $s_i = x$ for all players $i \in K$, we also write (x_K, s_{-K}) for s.

Given two joint strategies s' and s and a coalition K, we say that s' is a ***deviation of the players in*** K from s if $K = \{i \in N \mid s_i \neq s'_i\}$. We denote this by $s \xrightarrow{K} s'$. If in addition $p_i(s') > p_i(s)$ holds for all $i \in K$, we say that the deviation s' from s is ***profitable***. Further, we say that the players in K ***can profitably deviate from*** s if there exists a profitable deviation of these players from s.

Next, we call a joint strategy s a ***k-equilibrium***, where $k \in \{1, \dots, n\}$, if no coalition of at most k players can profitably deviate from s. Using this definition, a ***Nash equilibrium*** is a 1-equilibrium and a ***strong equilibrium***, see [5], is an n-equilibrium.

Given a joint strategy s, we call the sum

$$SW(s) = \sum_{i \in N} p_i(s)$$

the ***social welfare*** of s.

A ***coalitional improvement path***, in short a ***c-improvement path***, is a maximal sequence $\rho = (s^1, s^2, \dots)$ of joint strategies such that for every $k > 1$ there is a coalition K such that s^k is a profitable deviation of the players in K from s^{k-1}. If ρ is finite then by $last(\rho)$ we denote the last element of the sequence. Clearly, if a c-improvement path is finite, its last element is a strong equilibrium. We say that $\mathcal{G}$ has the ***finite c-improvement property*** (***c-FIP***) if every c-improvement path is finite. Further, we say that the function $P : S \to A$, where A is a set, is a ***generalized ordinal c-potential***, also called ***generalized strong potential***, see [7, 8], for $\mathcal{G}$ if for some strict partial ordering $(P(S), \succ)$ the fact that s' is a profitable deviation of the players in some coalition from s implies that $P(s') \succ P(s)$.

If a finite game admits a generalized ordinal c-potential then it has the c-FIP. The converse also holds, see, e.g., [3]. We say that $\mathcal{G}$ is ***c-weakly acyclic*** if for every joint strategy there exists a finite c-improvement path that starts at it. Note that games that have the c-FIP or are c-weakly acyclic game have a strong equilibrium.

We call a c-improvement path an ***improvement path*** if each deviating coalition consists of one player. The notions of a game having the ***FIP*** or being ***weakly acyclic***, see [15, 13], are then defined by referring to improvement paths instead of c-improvement paths.

3. COORDINATION GAMES ON DIRECTED GRAPHS

We now introduce the class of games we are interested in. Fix a finite set of colours M and a weighted directed graph (G, w) without self loops in which each edge e has a non-negative weight w_e associated with. We say that a node j is a ***neighbour*** of the node i if there is an edge $j \to i$ in G. Let N_i denote the set of all neighbours of node i in the graph G. By a ***colour assignment*** we mean a function that assigns to each node of G a finite non-empty set of colours. For technical reasons we also introduce the concept of a ***bonus***, which is a function β that to each node i and a colour c assigns a natural number $\beta(i, c)$. (We allow zero as a natural number.)

Given a weighted graph (G, w), a colour assignment A and a bonus function β we define a strategic game $\mathcal{G}(G, w, A, \beta)$ as follows:

- the players are the nodes,
- the set of strategies of player (node) i is the set of colours $A(i)$; we refer to the strategies as ***colours*** and to joint strategies as ***colourings***,
- each payoff function is defined by

$$p_i(s) = \sum_{j \in N_i,\, s_i = s_j} w_{j \to i} + \beta(i, s_i).$$

So each node simultaneously chooses a colour and the payoff to the node is the sum of the weights of the edges from its neighbours that chose its colour augmented by the bonus to the node from choosing the colour. We call these games ***coordination games on directed graphs***, from now on just ***coordination games***. Because weights are non-negative each coordination game satisfies the PPM.

In the paper we mostly consider the case when all weights are 1 and all bonuses are 0. Then each payoff function is simply defined by

$$p_i(s) := |\{j \in N_i \mid s_i = s_j\}|.$$

EXAMPLE 1. *Consider the directed graph and the colour assignment depicted in Figure 1.*

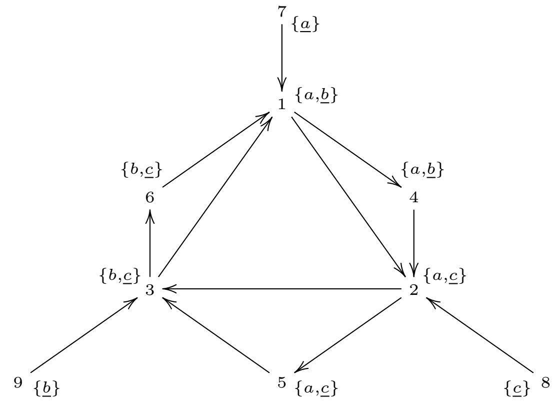


Figure 1: A directed graph with a colour assignment.

Take the joint strategy s that consists of the underlined strategies. Then the payoffs are as follows:

- *0 for the nodes 1, 7, 8 and 9,*
- *1 for the nodes 2, 4, 5, 6,*
- *2 for the node 3.*

Note that the above joint strategy is not a Nash equilibrium. For example, node 1 can profitably deviate to colour a. □

In what follows we study the problem of existence of Nash equilibria or strong equilibria in coordination games.

Finally, given a directed graph G and a set of nodes K, we denote by $G[K]$ the subgraph of G induced by K.

4. STRONG EQUILIBRIA

In this section we focus on the existence of strong equilibria. To start with, we have the following positive result.

Theorem 2. *Every coordination game whose underlying graph is a DAG has the c-FIP and a fortiori a strong equilibrium. Further, each Nash equilibrium is a strong equilibrium.*

Proof. Given a DAG $G := (V, E)$, where $V = \{1, \ldots, n\}$, we fix a permutation π of $1, \ldots, n$ such that for all $i, j \in V$

$$\text{if } i < j, \text{ then } (\pi(j) \to \pi(i)) \notin E. \tag{1}$$

So if $i < j$, then the payoff of the node $\pi(i)$ does not depend on the strategy selected by the node $\pi(j)$.

Then given a coordination game whose underlying directed graph is the DAG G we associate with each joint strategy s the sequence $p_{\pi(1)}(s), \ldots, p_{\pi(n)}(s)$ that we abbreviate to $p_\pi(s)$. We now claim that $p_\pi : S \to \mathbb{R}^n$ is a generalized ordinal c-potential when we take for the partial ordering $\succ$ on $p_\pi(S)$ the lexicographic ordering $>_{lex}$ on the sequences of reals.

Suppose that some coalition K profitably deviates from the joint strategy s to s'. Choose the smallest j such that $\pi(j) \in K$. Then $p_{\pi(j)}(s') > p_{\pi(j)}(s)$ and by (1) $p_{\pi(i)}(s') = p_{\pi(i)}(s)$ for $i < j$. This implies that $p_\pi(s') >_{lex} p_\pi(s)$, as desired. Hence the game has the c-FIP.

To prove the second claim, take a Nash equilibrium s and suppose it is not a strong equilibrium. Then some coalition K can profitably deviate from s to s'. Choose the smallest j such that $\pi(j) \in K$. Then $p_{\pi(j)}(s') > p_{\pi(j)}(s)$ and by (1) the payoff of $\pi(j)$ does not depend on the strategies selected by the other members of the coalition K. Hence $p_{\pi(j)}(s') = p_{\pi(j)}(s'_{\pi(j)}, s_{-\pi(j)})$, which contradicts the assumption that s is a Nash equilibrium. □

The next result deals with a class of coordination games introduced in [3]. Given the set of colours M, we say that a directed graph G is ***colour complete*** (with respect to a colour assignment A) if for every colour $x \in M$ each component of $G[V_x]$ is complete, where $V_x = \{i \in V \mid x \in A_i\}$. In particular, every complete graph is colour complete.

Theorem 3. *Every coordination game on a colour complete directed graph has the c-FIP and a fortiori a strong equilibrium.*

Proof. In [3] it is proved that every uniform game has the c-FIP, where we call a coordination game on a directed graph G ***uniform*** if for every joint strategy s and for every edge $i \to j \in E$ it holds: if $s_i = s_j$ then $p_i(s) = p_j(s)$. (In [3] only undirected graphs are considered, but the proof remains valid without any change.) Clearly every coordination game on a colour complete directed graph is uniform. □

It is difficult to come up with other classes of directed graphs for which the coordination game has the c-FIP. Indeed, consider the following example.

Example 4. *Consider a coordination game on a simple cycle $1 \to 2 \to \ldots \to n \to 1$, where $n \geq 3$ and such that the nodes share at least two colours, say a and b. Take the initial colouring $(a, b, \ldots, b)$. Then both $(a, \underline{b}, b, \ldots, b), (a, a, b, \ldots, b)$ and $(\underline{a}, a, b, \ldots, b), (b, a, b, \ldots, b)$ are profitable deviations. (To increase readability we underlined the strategies that were modified.) After these two steps we obtain a colouring that is a rotation of the first one. Iterating we obtain an infinite improvement path.*

Hence the coordination game does not have the FIP and a fortiori the c-FIP. □

However, a weaker result holds, which, for reasons that will soon become clear, we prove for a larger class of games.

Theorem 5. *Every coordination game with bonuses on a simple cycle is c-weakly acyclic, so a fortiori has a strong equilibrium.*

To prove it, we first establish a weaker claim.

Lemma 6. *Every coordination game with bonuses on a simple cycle is weakly acyclic.*

Proof. To fix the notation, suppose that the considered graph is $1 \to 2 \to \ldots \to n \to 1$. Below for $i \in \{2, \ldots, n\}$, $i \ominus 1 = i - 1$, and $1 \ominus 1 = n$.

Let $MA(i)$ be the set of available colours to player i with the maximal bonus, i.e., $MA(i) = \{c \in A(i) \mid \beta(i, c) = \max_{d \in A(i)} \beta(i, d)\}$. Let

$$BR(i, s_{-i}) = \{c \in MA(i) \mid \text{colour } c \text{ is a best response of player } i \text{ to } s_{-i}\}$$

be the set of best responses among the colours with the highest bonus only. The set $BR(i, s_{-i})$ is never empty because of the game structure and the fact that bonuses are natural numbers. Indeed, if $s_{i \ominus 1} \in MA(i)$, then $BR(i, s_{-i}) = \{s_{i \ominus 1}\}$ and otherwise $BR(i, s_{-i})$ is a non-empty subset of $MA(i)$.

Below we stipulate that whenever a player i updates in a joint strategy s his strategy to a best response to s_{-i}, he always selects a strategy from $BR(i, s_{-i})$.

Consider an initial joint strategy s. We construct a finite improvement path that starts with s as follows.

Phase 1. We proceed around the cycle and consider the players 1, 2, ..., $n-1$ in that order. For each player in turn, if his current strategy is not a best response, we update it to a best response respecting the above proviso. When this phase ends the current strategy of each of the players 1, 2, ..., $n-1$ is a best response.

If at this moment the current strategy of player n is also a best response, the current joint strategy s' is a Nash equilibrium and the path is constructed. Otherwise we move to the next phase.

Phase 2. We repeat the same process as in Phase 1, but starting with s' and player n and proceeding at most n steps. From now on at each step at least $n-1$ players have a best response strategy. So if at a certain moment the current strategy of the considered player is a best response, the current joint strategy is a Nash equilibrium and the path is constructed. Otherwise, after n steps, we move to the final phase.

Phase 3. We repeat the same process as in Phase 2. Now in the initial joint strategy each player i has a strategy from $MA(i)$. Because of the definition of $BR(i, s_{-i})$ each player can improve his payoff only if he switches to the strategy selected by his predecessor. So after at most n steps this phase terminates and we obtain a Nash equilibrium. □

By Lemma 6 every coordination game on a simple cycle has a Nash equilibrium. However, not every Nash equilibrium is then a strong equilibrium.

Example 7. *Consider the directed graph depicted in Figure 2, together with the sets of colours associated with the nodes.*

Clearly (a, b) is a Nash equilibrium. However, it is not a strong equilibrium since the coalition $\{1, 2\}$ can profitably deviate to (c, c), which is a strong equilibrium. □

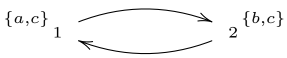


Figure 2: Nash equilibria versus strong equilibria

On the other hand, the following observation holds.

Lemma 8. *Consider a coordination game with bonuses on a simple cycle with n nodes. Then every Nash equilibrium is a k-equilibrium for all $k \in \{1, \ldots, n-1\}$.*

Proof. Take a Nash equilibrium s. It suffices to prove that it is an $(n-1)$-equilibrium. Suppose otherwise. Then for some coalition K of size $\leq n-1$ and a joint strategy s', $s \xrightarrow{K} s'$ is a profitable deviation.

Assume $k \ominus 1 = k - 1$ if $k > 1$ and $1 \ominus 1 = n$. Take some $i \in K$ such that $i \ominus 1 \notin K$. We have $p_i(s') > p_i(s)$. Also $p_i(s'_i, s_{-i}) = p_i(s')$, since $s_{i\ominus 1} = s'_{i\ominus 1}$. So $p_i(s'_i, s_{-i}) > p_i(s)$, which contradicts the fact that s is a Nash equilibrium. □

Proof of Theorem 5. Take a joint strategy s. By Lemma 6 a finite improvement path starts at s and ends in a Nash equilibrium s'. By Lemma 8 s' is an $(n-1)$-equilibrium. If s' is not a strong equilibrium, then a profitable deviation $s' \xrightarrow{N} s''$ exists, where, recall, N is the set of all players. Because of the game structure the social welfare along each c-improvement path weakly increases, while in the last step the social welfare strictly increases. So $SW(s'') > SW(s)$.

If s'' is not a strong equilibrium, we repeat the above procedure starting with s''. Since each time the social welfare strictly increases, eventually this process stops and we obtain a finite c-improvement path. □

Using Theorem 5, we now show that every coordination game in which all strongly connected components are simple cycles is c-weakly acyclic. We first introduce some notations and make use of the following well-known decomposition result.

Theorem 9 ([6], page 92). *Every directed graph is a directed acyclic graph (DAG) of its strongly connected components (SCCs).*

Given a graph $G = (V, E)$, let $D = (V_D, E_D)$ be the corresponding DAG obtained by the above decomposition theorem and let $g : 2^V \to V_D$ be the function that maps each SCC in G to a node in D. Let $g^{-1}(v) = X \subseteq V$ where $g(X) = v$. Note that for each $i \in V$, there is a unique $v \in V_D$ such that $i \in g^{-1}(v)$, we denote this node by v_i. Let $|V_D| = m$ and $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$ be a topological ordering of V_D (this is well-defined since D is a DAG). We define a labelling function $l_D : V_D \to \{1, \ldots, m\}$ as follows: for all $v \in V_D$, $l_D(v) = j$ iff $\theta_j = v$. We can extend l_D to a function $l : V \to \{1, \ldots, m\}$ in the natural way: for all $i \in V$, $l(i) = l_D(v)$ if $i \in g^{-1}(v)$.

Note that for each node $v \in V_D$, either $g^{-1}(v) = \{i\}$ for some $i \in V$ or $g^{-1}(v) = X \subseteq V$ with $|X| \geq 2$ and the subgraph of G induced by the set of nodes X forms an SCC. Also, note that every $v \in V_D$ and a joint strategy s in $\mathcal{G}$, defines a coordination game with bonuses $\mathcal{G}_v$ on the graph $G(v, s) = (V', E')$ which is the subgraph induced by the set of nodes $V' = g^{-1}(v)$. For $i \in V'$ and $a \in A(i)$ we put $\beta(i, a) := |\{j \in N_i \setminus V' \mid s_j = a\}|$.

Theorem 10. *Every coordination game on a directed graph G in which all strongly connected components of G are simple cycles is c-weakly acyclic and a fortiori has a strong equilibrium.*

Proof. Consider a coordination game $\mathcal{G}$ on a graph $G = (V, E)$ where all SCCs are simple cycles. Let $D = (V_D, E_D)$ be the corresponding DAG with $|V_D| = m$. Since all SCCs in G are simple cycles, it follows that for all $v \in V_D$, either $g^{-1}(v) = \{i\}$ or $g^{-1}(v) = X \subseteq V$ such that the induced graph on X forms a simple cycle in G.

Let $v \in V_D$ such that the induced graph on $g^{-1}(v)$ forms a simple cycle in G. For a joint strategy t in $\mathcal{G}$, consider the resulting game $\mathcal{G}_v$ on the graph (V', E'), the subgraph induced by the set of nodes $g^{-1}(v)$. Let $s^0 = t_{V'}$ (the restriction of the joint strategy t to nodes in V') and let $\rho : s^0, s^1, \ldots, s^k$ be a finite c-improvement path in $\mathcal{G}_v$ which is guaranteed to exist by Theorem 5. Define $\mathit{CPath}(\mathcal{G}_v, t)$ as follows:

$$\mathit{CPath}(\mathcal{G}_v, t) = \begin{cases} \epsilon & \text{if } t_{V'} \text{ is a strong equilibrium in } \mathcal{G}_v, \\ \lambda_t(s^1), \ldots, \lambda_t(s^k) & \text{otherwise,} \end{cases}$$

where for all $h \in \{1, \ldots, k\}$, $\lambda_t(s^h)$ is the joint strategy in $\mathcal{G}$ defined as: for all $i \in V$, $(\lambda_t(s^h))_i = s^h_i$ if $i \in V'$ and $(\lambda_t(s^h))_i = t_i$ if $i \notin V'$.

For a joint strategy t in $\mathcal{G}$ and $v \in V_D$, if the underlying graph of the coordination game $\mathcal{G}_v$ with bonuses consists of exactly one node, then the game is trivially c-weakly acyclic and we define $\mathit{CPath}(\mathcal{G}_v, t)$ analogously.

Let t^0 be an arbitrary joint strategy in $\mathcal{G}$. We define a sequence of sequences of joint strategies as follows:

- $\rho_0 = t^0$,
- for $h \in \{0, 1, \ldots, m-1\}$, let $\rho_{h+1} = \rho_h \cdot \mathit{CPath}(\mathcal{G}_v, t^h)$ where $l_D(v) = h + 1$ and $t^h = \mathit{last}(\rho_h)$.

Let $\rho = \rho_m$. From the definition of ρ_m and CPath, it follows that ρ is a finite sequence of joint strategies in $\mathcal{G}$. By induction on the length of ρ, we can show that for every subsequent pair of joint strategies t^k and t^{k+1} in ρ, there is a coalition $K \subseteq V$ for which t^{k+1} is a profitable deviation from t^k. To complete the proof, it suffices to argue that ρ is maximal, or equivalently, that $\mathit{last}(\rho)$ is a strong equilibrium.

Suppose $\mathit{last}(\rho)$ is not a strong equilibrium. Then there exists $K \subseteq V$ and a joint strategy s such that there is a profitable deviation of players in K from $\mathit{last}(\rho)$ to s. Let d be the least element of the set $\{l(i) \mid i \in K\}$ and $X = K \cap \{i \in V \mid l(i) = d\}$. By definition of a profitable deviation, we have that for all $i \in X$, $p_i(s) > p_i(\mathit{last}(\rho))$. Note that for all $i \in X$ and for all $j \in N_i \setminus g^{-1}(v_i)$, we have $l(j) < d$. Therefore, $(N_i \setminus g^{-1}(v_i)) \cap K = \emptyset$. Also note that for all $j \in g^{-1}(v_i)$, $(\mathit{last}(\rho_d))_j = (\mathit{last}(\rho))_j$. But this implies that the coalition X has a profitable deviation from the joint strategy $(\mathit{last}(\rho_d))_X$ to s_X in the game $\mathcal{G}_{v_i}$. This contradicts the fact that $\mathit{last}(\rho_d)$ is a strong equilibrium in the game $\mathcal{G}_{v_i}$. □

We conclude this section by considering another class of coordination games. Example 4 shows that even when only two colours are used, the coordination game does not need to have the c-FIP. On the other hand, a weaker property does hold.

Theorem 11. *Every coordination game in which only two colours are used is c-weakly acyclic and a fortiori has a strong equilibrium.*

Proof. We prove the result for a more general class of games, namely the ones that satisfy the PPM. Call the colours blue and red, that we abbreviate to b and r. When a node selected blue we refer to it as a blue node, and the same for the red colour.

Take a joint strategy s^1. Consider a maximal sequence ξ of profitable deviations of the coalitions starting in s in which the nodes can only switch to blue. At each step the number of blue nodes increases, so ξ is finite. Let $s^1, \ldots, s^k$, where $k \geq 1$, be the successive joint strategies of ξ.

If s^k is a strong equilibrium, then ξ is the desired finite improvement path. Otherwise consider a maximal sequence χ of profitable deviations of the coalitions starting in s^k in which the nodes can only switch to red. χ is finite. Let $s^k, s^{k+1}, \ldots, s^{k+l}$, where $l \geq 1$, be the successive joint strategies of χ.

We claim that s^{k+l} is a strong equilibrium. Suppose otherwise. Then for some joint strategy s', $s^{k+l} \xrightarrow{K} s'$ is a profitable deviation of some coalition K. Let L be the set of nodes from K that switched in this deviation to blue. By the definition of s^{k+l} the set L is non-empty.

Given a set of nodes M and a joint strategy s we denote by $(M : b, s_{-M})$ the joint strategy obtained from s by letting the nodes in M to select blue, and similarly for the red colour. Also it should be clear what joint strategy we denote by $(M : b, P \setminus M : r, s_{-P})$, where $M \subseteq P$.

We claim that $s^{k+l} \xrightarrow{L} (L : b, s^{k+l}_{-L})$ is a profitable deviation of the players in L. Indeed, we have for all $i \in L$

$$p_i(L : b, s^{k+l}_{-L}) > p_i(s^{k+l}), \tag{2}$$

since by the PPM $p_i(L : b, s^{k+l}_{-L}) \geq p_i(s')$ and by the assumption $p_i(s') > p_i(s^{k+l})$.

Let M be the set of nodes from L that are red in s^k. Suppose that M is non-empty. We show that then

$$p_M(M : r, L \setminus M : b, s^k_{-L}) < p_M(M : b, L \setminus M : b, s^k_{-L}). \tag{3}$$

Indeed, we have for all $i \in M$

$$\begin{aligned} & p_i(M : r, L \setminus M : b, s^k_{-L}) \leq p_i(M : r, L \setminus M : b, s^{k+l}_{-L}) \\ \leq \;& p_i(M : r, L \setminus M : r, s^{k+l}_{-L}) < p_i(M : b, L \setminus M : b, s^{k+l}_{-L}) \\ \leq \;& p_i(M : b, L \setminus M : b, s^k_{-L}), \end{aligned}$$

where the weak inequalities are due to the PPM and the strict inequality holds by the definition of L.

But $s^k = (M : r, L \setminus M : b, s^k_{-L})$, so (3) contradicts the definition of s^k. So M is empty, i.e., all nodes from L are blue in s^k. We now have for all $i \in L$

$$\begin{aligned} & p_i(L : r, s^k_{-L}) \leq p_i(L : r, s^{k+l}_{-L}) = p_i(s^{k+l}) \\ < \;& p_i(L : b, s^{k+l}_{-L}) \leq p_i(L : b, s^k_{-L}), \end{aligned}$$

where again the weak inequalities are due to the PPM and the strict inequality holds by (2).

But $(L : r, s^k_{-L}) = s^k$, so we proved that $s^k \xrightarrow{L} (L : b, s^k_{-L})$ is a profitable deviation. This yields a contradiction with the definition of s^k. □

The following example shows that when three colours are used, Nash equilibria, so a fortiori strong equilibria do not need to exist.

Example 12. *Consider the directed graph depicted in Figure 1 of Example 1, together with the sets of colours associated with the nodes. We argue that the coordination game associated with this graph does not have a Nash equilibrium. Note that for nodes 7, 8 and 9 the only option is to select the unique strategy in its strategy set. The best response for nodes 4, 5 and 6 is to always select the same strategy as nodes 1, 2 and 3 respectively. Therefore, to show that the game does not have a Nash equilibrium, it suffices to consider the strategies of nodes 1, 2 and 3. We denote this by the triple (s_1, s_2, s_3). Below we list all such joint strategies and we underline a strategy that is not a best response to the choice of other players: $(\underline{a}, a, b)$, $(a, a, \underline{c})$, $(a, c, \underline{b})$, $(a, \underline{c}, c)$, $(b, \underline{a}, b)$, $(\underline{b}, a, c)$, $(b, c, \underline{b})$ and $(\underline{b}, c, c)$.* □

Call now a graph a ***coloured DAG*** (with respect to a colour assignment A) if for each available colour x the components of the subgraph induced by the nodes having colour x are DAGs. In view of Theorem 3 it is tempting to try to generalize Theorem 2 to coloured DAGs. However, the directed graph depicted in Figure 1 is a coloured DAG and, as explained in the above example, the coordination game on this graph has no Nash equilibrium.

5. COMPLEXITY ISSUES

Next, we study the complexity of the existence problems and of the problem of finding strong equilibria.

Theorem 13. *The Nash equilibrium existence problem in coordination games is NP-complete.*

Proof. The problem is in NP, since we can simply guess a colour assignment and checking whether it is a Nash equilibrium can be done in polynomial time.

To prove NP-hardness we provide a reduction from the 3-SAT problem, which is NP-complete. Notice that an edge with a natural number weight w can be simulated by adding w extra players to the game. More precisely, an edge $(i \to j)$ with the weight w can be simulated by the extra set of players $\{i_1, \ldots, i_w\}$ and the following $2 \cdot w$ unweighted edges: $\{(i \to i_1), (i \to i_2), \ldots, (i \to i_w), (i_1 \to j), (i_2 \to j), \ldots, (i_w \to j)\}$. Given a colour assignment in the game with the weighted edges, we then assign to each of the nodes $i_1, \ldots, i_w$ the colour set of the node i.

Therefore we will assume that edges can have such weights assigned to them, because this simplifies our construction. Assume we are given a 3-SAT formula

$$\phi = (a_1 \vee b_1 \vee c_1) \wedge (a_2 \vee b_2 \vee c_2) \wedge \ldots \wedge (a_k \vee b_k \vee c_k)$$

with k clauses and n propositional variables $x_1, \ldots, x_n$, where each a_i, b_i, c_i is a literal equal to x_j or $\neg x_j$ for some j. We will construct a coordination game $\mathcal{G}_\phi$ of size $\mathcal{O}(k)$ such that $\mathcal{G}_\phi$ has a Nash equilibrium iff ϕ is satisfiable.

First, for every propositional variable x_i we have a corresponding node X_i in $\mathcal{G}_\phi$ with two possible colours $\top$ and $\bot$. Intuitively, for a given truth assignment, if x_i is true then $\top$ should be chosen for X_i and otherwise $\bot$ should be chosen. In our construction we make use of the following gadget, denoted by $D_i(x, y, z)$, with three parameters $x, y, z \in \{\top, \bot\}$ and i used just for labelling purposes, and presented in Figure 3. This gadget behaves similarly to the game without Nash equilibrium analyzed in Example 12.

What is important is that for all possible parameters values, the gadget $D_i(x, y, z)$ does not have a Nash equilibrium. Indeed, each of the nodes A_i, B_i, or C_i can always secure a payoff 2, so selecting $\top$ or $\bot$ is never a best response and

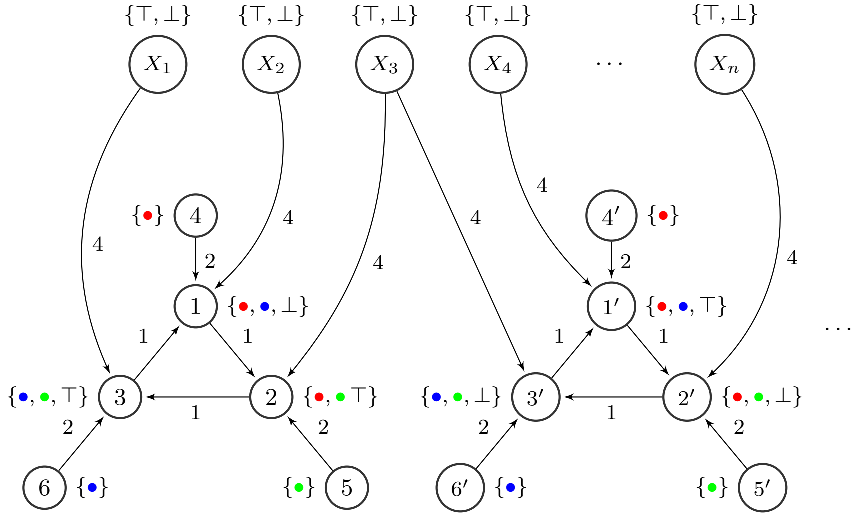


Figure 4: The game $\mathcal{G}_\phi$ corresponding to the formula $\phi = (x_1 \vee \neg x_2 \vee x_3) \wedge (\neg x_3 \vee x_4 \vee \neg x_n)$, where in each gadget the nodes of indegree 1 are omitted.

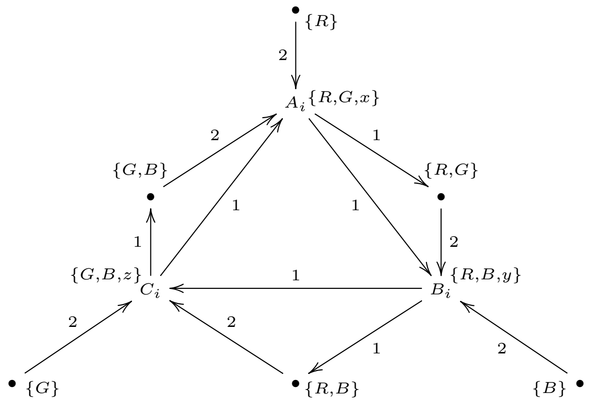


Figure 3: Gadget D_i with three parameters $x, y, z \in \{\top, \bot\}$ and three distinguished nodes A_i, B_i, C_i.

hence in no Nash equilibrium a node chooses $\top$ or $\bot$. The rest of the reasoning is as in Example 12.

For any literal, l, let

$$\mathrm{pos}(l) := \begin{cases} \top & \text{if } l \text{ is a positive literal} \\ \bot & \text{otherwise} \end{cases}$$

For every clause $(a_i \vee b_i \vee c_i)$ in ϕ we add to the game graph $\mathcal{G}_\phi$ the $D_i(\mathrm{pos}(a_i), \mathrm{pos}(b_i), \mathrm{pos}(c_i))$ instance of the gadget. Finally, for every literal a_i, b_i, or c_i in ϕ, which is equal to x_j or $\neg x_j$ for some j, we add an edge from X_j to A_i, B_i, or C_i, respectively, with weight 4. We depict an example game $\mathcal{G}_\phi$ in Figure 4.

We claim that $\mathcal{G}_\phi$ has a Nash equilibrium iff ϕ is satisfiable.

($\Rightarrow$) Assume there is a Nash equilibrium s in the game $\mathcal{G}_\phi$. We claim that the truth assignment $\nu : \{x_1, \ldots, x_n\} \to \{\top, \bot\}$ that assigns to each x_j the colour selected by the node X_j in s makes ϕ true. Fix $i \in \{1, \ldots, k\}$. We need to show that ν makes one of the literals a_i, b_i, c_i of the clause $(a_i \vee b_i \vee c_i)$ true.

From the above observation about the gadgets it follows that at least one of the nodes A_i, B_i, C_i selected in s the same colour as its neighbour X_j. Without loss of generality suppose it is A_i. The only colour these two nodes, A_i and X_j, have in common is $\mathrm{pos}(a_i)$. So X_j selected in s $\mathrm{pos}(a_i)$, which by the definition of ν equals $\nu(x_j)$. Moreover, by construction x_j is the variable of the literal a_i. But $\nu(x_j) = \mathrm{pos}(a_i)$ implies that ν makes a_i true.

($\Leftarrow$) Assume ϕ is satisfiable. Take a truth assignment $\nu : \{x_1, \ldots, x_n\} \to \{\top, \bot\}$ that makes ϕ true. For all j, we assign to the node X_j the colour $\nu(x_j)$. We claim that this assignment can be extended to a Nash equilibrium in $\mathcal{G}_\phi$.

Fix $i \in \{1, \ldots, k\}$ and consider the $D_i(\mathrm{pos}(a_i), \mathrm{pos}(b_i), \mathrm{pos}(c_i))$ instance of the gadget. The truth assignment ν makes the clause $(a_i \vee b_i \vee c_i)$ true. Suppose without loss of generality that ν makes a_i true. We claim that then it is always a unique best response for the node A_i to select the colour $\mathrm{pos}(a_i)$.

Indeed, let j be such that $a_i = x_j$ or $a_i = \neg x_j$. Notice that the fact that ν makes a_i true implies that $\nu(x_j) = \mathrm{pos}(a_i)$. So when node A_i selects $\mathrm{pos}(a_i)$, the colour assigned to X_j, its payoff is 4.

This partial assignment of colours can be completed to a Nash equilibrium. Indeed, remove from the directed graph of $\mathcal{G}_\phi$ all X_j nodes and the nodes that secured the payoff 4, together with the edges that use any of these nodes. The resulting graph has no cycles, so by Theorem 2 the corresponding coordination game has a Nash equilibrium. Combining both assignments of colours we obtain a Nash equilibrium in $\mathcal{G}_\phi$. □

Corollary 14. *The strong equilibrium existence problem in coordination games is NP-complete.*

Proof. It suffices to note that in the above proof the ($\Rightarrow$) implication holds for a strong equilibrium, as well, while in the proof of the ($\Leftarrow$) implication by virtue of Theorem 2 actually a strong equilibrium is constructed. □

An interesting application of Theorem 13 is in the context of polymatrix games. These are finite strategic form games in which the influence of a pure strategy selected by any player on the payoff of any other player is always the same, regardless of what strategies other players select. Formally, for all pairs of players i and j there exists a partial payoff function a^{ij} such that for any joint strategy $s = (s_1, \ldots, s_n)$, the payoff of player i is given by $p_i(s) := \sum_{j \neq i} a^{ij}(s_i, s_j)$. In [14] we proved that deciding whether a polymatrix game has a Nash equilibrium is NP-complete. We can strengthen this result as follows.

Theorem 15. *Deciding whether a polymatrix game with 0/1 partial payoffs has a Nash equilibrium is NP-complete.*

Proof. We can efficiently translate any coordination game $\mathcal{G}(G, M, w, A, \beta)$ into a polymatrix game $\mathcal{P}$ with only 0/1 partial payoff as follows. The number of players in $\mathcal{P}$ will be equal to the number of nodes in G and the set of strategies for each player will be M. We define $a^{ij}(s_i, s_j) := 1$ if $s_i = s_j$ and $j \in N_i$, and $a^{ij}(s_i, s_j) := 0$ otherwise.

Notice that any joint strategy $s = (s_1, \ldots, s_n)$ in $\mathcal{G}$ is also a joint strategy in $\mathcal{M}$ with exactly the same payoff, because $p_i^{\mathcal{P}}(s) = \sum_{j \neq i} a^{ij}(s_i, s_j) = |\{j \in N_i \mid s_i = s_j\}| = p_i^{\mathcal{G}}(s)$. It follows that Nash equilibria in $\mathcal{G}$ and $\mathcal{P}$ coincide. In particular, there exists Nash equilibrium in $\mathcal{G}$ if and only if there exists one in $\mathcal{P}$, but the former problem was shown to be NP-hard in Theorem 13, so the latter is also NP-hard. On the other hand, for any polymatrix game we can guess a joint strategy and check whether it is a Nash equilibrium in polynomial time, which shows this decision problem is in fact NP-complete. □

Next, we determine the complexity of finding a strong equilibrium. We begin with the following auxiliary result.

Theorem 16. *A strong equilibrium of a coordination game with bonuses on a simple cycle can be found in linear time.*

Proof. Let n be the number of players in the game and C the number of possible colours. We assume adjacency list representation for the game graph, binary representation of the bonuses and that the list of colours available to player i is given as a list of length $|A(i)|$ of elements of size $\log C$. Formally, the size of the input for player i only is $\Theta(|A(i)| \log C + \sum_{c \in A(i)} \log(\beta(i, c) + 1))$; the sum of these over $i = 1, \ldots, n$ gives the total size of the input.

First note that for any colour assignment, the best response of the i-th player can be found in time linear in the size of her part of the input just by checking all possible colours in $A(i)$. Second, each phase of the algorithm in Lemma 6 looks for the best response (with a preference given to colours with a higher bonus) of each player at most once, which will require time linear in the size of the whole input. The algorithm requires at most three such phases before a Nash equilibrium is found, so it runs in linear time.

Note that thanks to Lemma 8 we know that any NE in such a game structure is already a $(n-1)$-equilibrium, so the only way this joint strategy is not a strong equilibrium is when all n players can strictly improve their payoff. However, in any Nash equilibrium a player has to have her payoff at at most one below the maximum possible one, because that is the minimum payoff for picking a colour with the highest bonus. Moreover, player's payoff can only be a natural number.

Therefore, the only possibility when a NE is not a strong equilibrium is when there is a joint strategy which gives all the players their maximum possible payoff, i.e. each player is assigned a colour with the highest possible bonus as well as gets an extra +1 to her payoff for colour agreement with her only neighbour. The latter implies that all the players need to pick the same colour in such a joint strategy.

To check whether such a joint strategy exists we do the following. Let $p = \mathrm{argmin}_i |A(i)|$ be the player with the least number of colours to choose from. We pick the set of her colours with the maximal bonus and intersect it with the set of colours with the maximal bonus for every other player. An intersection of two sets represented as lists of length a and b of elements of size K can be done in $\Theta(aK + bK)$ time, so the total running time will be $\Theta(n|A(p)| \log C + \sum_{i=1}^{n} |A(i)| \log C) = \Theta(\sum_{i=1}^{n} |A(i)| \log C)$, because $|A(p)| \leq |A(i)|$ for all i, which is linear. If the final set is empty then any NE is a strong equilibrium and otherwise we know how to construct one. □

Corollary 17. *A strong equilibrium of a coordination game on a graph in which all strongly connected components are simple cycles can be computed in linear time.* □

6. CONCLUSIONS

We presented here a study of a simple class of coordination games on directed graphs. We focused on the existence of Nash and strong equilibria. We also studied the complexity of checking for the existence of Nash and strong equilibria, as well as the complexity of computing a strong equilibrium in certain cases where it is guaranteed to exist.

A number of open problems remain. We showed that in general Nash equilibria and strong equilibria are not guaranteed to exist. However, if the underlying graph is a DAG, is colour complete or is such that every SCC is a simple cycle, then strong equilibria always exist. It would be interesting to identify other classes of graphs for which Nash or strong equilibria exist.

The proof of Lemma 6 shows that in the case of a simple cycle, starting from any initial joint strategy a Nash equilibrium can be found by an improvement path of length at most $3n$. Also, each step of such a path can be constructed in linear time. Additionally, the proof of Theorem 5 shows that a strong equilibrium can be found by an improvement path of length at most $3n + 1$, possibly augmented by a single profitable deviation of all players. It would be interesting to extend this analysis of bounds on the lengths of improvement paths to other cases when a Nash or a strong equilibrium is known to exist.

In the future we plan to study the inefficiency of equilibria in coordination games on directed graphs. Also, we plan to study coordination games on finite directed weighted graphs. While we already defined here these games, we used weights solely as a means to simplify the argument in the proof of Theorem 13. It should be noted that Lemma 6 does not hold for finite directed weighted graphs and, as a consequence, Theorems 5, 10, and 16 do not hold either. A counterexample to Lemma 6 can be constructed by modifying the game in Figure 1 as follows. Nodes 4, 5, 6 are removed and replaced by assigning weight 2 to all the edges in the cycle. Nodes 7, 8, 9 are also removed and replaced by a +1 bonus to the colour of the node removed. It is easy to see that the behaviour of this new game will mimic the game in Figure 1. On the other hand, Theorem 2 and its proof is still valid for finite directed weighted graphs as well is Theorem 13, because checking whether a colour assignment is a Nash equilibrium can still be done in polynomial time for them.

As an example of coordination games on weighted directed graphs consider the problem of a choice of the trade treaties between various countries. Assume a directed weighted graph in which the nodes are the countries and the weight on an edge $i \to j$ corresponds to the percentage of the overall import of country j from country i. Suppose additionally that each country should choose a specific trade treaty, that the options for the countries differ (for instance because of its geographic location) and that each treaty offers the same tax-free advantages. Then once the countries choose the treaties, the payoff to each country is the aggregate percentage of its import that is tax-free.

The case of weighted directed graphs can be seen as a minor modification of the ***social network games*** with obligatory product selection that we introduced and analyzed in [4]. These are games associated with a threshold model of a social network introduced in [1] which is based on weighted graphs with thresholds. The difference consists of using thresholds equal to 0. However, setting the thresholds to 0 essentially changes the nature of the games and crucially affects the validity of several arguments.

Acknowledgments

We are grateful to Mona Rahn and Guido Schäfer for useful discussions and thank Piotr Sankowski and the referees for helpful comments. First author is also a Visiting Professor at the University of Warsaw. He was partially supported by the NCN grant nr 2014/13/B/ST6/01807.

7. REFERENCES

[1] K. R. Apt and E. Markakis. Diffusion in social networks with competing products. In *Proc. 4th International Symposium on Algorithmic Game Theory (SAGT11)*, volume 6982 of *Lecture Notes in Computer Science*, pages 212–223. Springer, 2011.

[2] K. R. Apt, M. Rahn, G. Schäfer, and S. Simon. Coordination games on graphs (extended abstract). In *Proc. 10th International Workshop on Internet and Network Economics (WINE)*, volume 8877 of *Lecture Notes in Computer Science*, pages 441–446. Springer, 2014.

[3] K. R. Apt, M. Rahn, G. Schäfer, and S. Simon. Coordination games on graphs, 2015. Available from `http://arxiv.org/abs/1501.07388`.

[4] K. R. Apt and S. Simon. Social network games with obligatory product selection. In *Proc. 4th International Symposium on Games, Automata, Logics and Formal Verification (GandALF '13)*, volume 119, pages 180–193. EPTCS, 2013.

[5] R. J. Aumann. Acceptable points in general cooperative n-person games. In R. D. Luce and A. W. Tucker, editors, *Contribution to the theory of game IV, Annals of Mathematical Study 40*, pages 287–324. University Press, 1959.

[6] S. Dasgupta, C. Papadimitriou, and U. Vazirani. *Algorithms.* McGraw-Hill, 2006.

[7] T. Harks, M. Klimm, and R. H. Möhring. Strong equilibria in games with the lexicographical improvement property. *International Journal of Game Theory*, 42(2):461–482, 2013.

[8] R. Holzman and N. Law-Yone. Strong equilibrium in congestion games. *Games and Economic Behavior*, 21(1-2):85–101, 1997.

[9] M. Jackson and Y. Zenou. Games on networks. In H. P. Young and S. Zamir, editors, *Handbook of Game Theory 4*, pages 95–163. Elsevier, 2014.

[10] E. Janovskaya. Equilibrium points in polymatrix games. *Litovskii Matematicheskii Sbornik*, 8:381–384, 1968.

[11] M. Kearns, M. Littman, and S. Singh. Graphical models for game theory. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (UAI '01)*, pages 253–260. Morgan Kaufmann, 2001.

[12] H. Konishi, M. Le Breton, and S. Weber. Equivalence of strong and coalition-proof Nash equilibria in games without spillovers. *Economic Theory*, (9):97–113, 1997.

[13] I. Milchtaich. Congestion games with player-specific payoff functions. *Games and Economic Behaviour*, 13:111–124, 1996.

[14] S. Simon and K. R. Apt. Social network games. *Journal of Logic and Computation*, 1(25):207–242, 2015.

[15] H. P. Young. The evolution of conventions. *Econometrica*, 61(1):57–84, 1993.

On the Solvability of Inductive Problems: A Study in Epistemic Topology

Alexandru Baltag
Institute for Logic, Language
and Computation
Science Park 107
Amsterdam, The Netherlands
a.baltag@uva.nl

Nina Gierasimczuk
Institute for Logic, Language
and Computation
Science Park 107
Amsterdam, The Netherlands
n.gierasimczuk@uva.nl

Sonja Smets
Institute for Logic, Language
and Computation
Science Park 107
Amsterdam, The Netherlands
S.J.L.Smets@uva.nl

ABSTRACT

We investigate the issues of inductive problem-solving and learning by doxastic agents. We provide topological characterizations of solvability and learnability, and we use them to prove that AGM-style belief revision is "universal", i.e., that every solvable problem is solvable by AGM conditioning.

Keywords

Learning Theory, Topological Epistemology, Belief Revision

1. INTRODUCTION

When in the course of observations it becomes necessary for agents to arrive at a generalization, they should declare, along with their conjecture, the extent of their certainty. The problem of induction seems formidable if a standard of absolute certainty is imposed on the learner. Indeed, as is well-known in Philosophy of Science, the so-called problem of empirical underdetermination (i.e., the fact that typically the data are compatible with more than one hypothesis) rules out any chance of obtaining infallible knowledge in empirical research. But apart from the conclusions based on absolute certainty (cf. [13, 10, 15]), learners can produce hypotheses based on *beliefs*. It is thus strange that Formal Learning Theory and Belief Revision Theory developed completely independently from each other, and that they have generally maintained their distance ever since.

However, there does exist a line of research that combines belief revision with learning-theoretic notions, line pursued by Kelly [21, 20], Kelly, Schulte and Hendricks [26], Martin and Osherson [28], Osherson [29] and ourselves [13, 3, 4, 14]. In this paper we continue this research program, using topological characterizations and methods.

An *inductive problem* consists of a state space, a family of "potential observations", and a "question" (i.e., a partition of the state space). These observations provide data for learning. The problem is *solvable* if there exists a learner that, after observing "enough" pieces of data, eventually stabilizes on the correct answer. A special case of solvability is *learnability in the limit*, corresponding to the solvability of the "ultimate" question: 'What is the actual state of the world?'. This notion matches the usual learning-theoretic concept of *identifiability in the limit* [32, 16, 30].

The aim of the paper is twofold. First, we give topological characterizations of the notions of solvability (and learnability), in terms of topological separation principles. Intuitively, the ability to reliably learn the true answer to a question, is related to the ability to "separate" answers by observations. The second goal is to use these topological results to look at the "solving power" of well-behaved doxastic agents, such as the ones whose beliefs satisfy the usual $KD45$ postulates of doxastic logic, as well as the standard AGM postulates of rational belief-revision [1]. We look at a particularly simple and canonical type of doxastic agent, who forms beliefs by *AGM conditioning*.

Our main result is that AGM conditioning is *universal for problem-solving*, i.e., that every solvable problem can be solved by AGM conditioning. This means that (contrary to some prior claims), AGM belief-revision postulates are not an obstacle to problem-solving. As a special case, it follows that AGM conditioning is also "universal for learning" (every learnable space can be learned by conditioning).[1]

The close connections between Epistemology and General Topology have already been noticed long ago [33, 19]. Based on these connections, Kevin Kelly started a far-reaching program [19, 22] meant to import ideas and techniques from both Formal Learning Theory and Topology into mainstream Epistemology, and show their relevance to the induction problem in Philosophy of Science. A further connection is the one with Ockham's Razor, that would

> *(...) guarantee that always choosing the simplest theory compatible with experience and hanging on to it while it remains the simplest is both necessary and sufficient for efficiency of inquiry.* [22]

Simplicity has been claimed to have topological characteristics—the simplicity order should in some way follow the structure imposed on the uncertainty range by possible tests and observations. It has also been linked with the notion of minimal mind change, where the learning agent keeps the conjecture changes to a minimum [19, 31].

Taken together, our results can be seen as a vindication both of the general topological program in Inductive Epistemology [19, 22] and of the AGM Belief Revision Theory [1]. On the first front, our general topological characterizations of learning-theoretic concepts seem to confirm Kelly's long-standing claim that Inductive Epistemology can be seen mathematically as a branch of General Topology. On the second front, our universality result seems to vindicate Be-

[1] This special case is a topological translation of one of our previous results [3, 4]. However, the result about problem-solving universality is not only new and much more general, but also much harder to prove, involving new topological notions and results.

lief Revision Theory as a canonical form of learning.[2]

2. EPISTEMIC SPACES AND INDUCTIVE PROBLEMS

Definition 1. *An* epistemic space *is a pair $\mathbb{S} = (S, \mathcal{O})$ consisting of a state space S and a countable (or finite) set of* observable properties *("data") $\mathcal{O} \subseteq \mathcal{P}(S)$. We denote by by $\mathcal{O}_s := \{O \in \mathcal{O} \mid s \in O\}$ the set of all observable properties (holding) at a given state s.*

One can think of the states in S as "possible worlds", in the tradition of Kripke and Lewis.The sets $O \in \mathcal{O}$ represent properties of the world that are in principle observable: if true, such a property will eventually be observed (although there is no upper bound on the time needed to come to observe it).

To keep things simple, we assume that at each step of the learning process only one property is observed. As for the countability of the set $\mathcal{O}$, it is natural to think of observables as properties which can be expressed by means of a language or numerical coding system, generated from a grammar with a finite vocabulary. Any such family $\mathcal{O}$ will be (at most) countable.

We denote by $\mathcal{O}^{\cap}$ the family of all finite intersections of observations from $\mathcal{O}$, and by $\mathcal{O}^*$ the family of all finite sequences of observations. Such a finite sequence $\sigma = (O_0, O_1, \ldots, O_i) \in \mathcal{O}^*$ is called a *data sequence*, and its i-th component is denoted by $\sigma_i := O_i$. It is easy to see that both $\mathcal{O}^{\cap}$ and $\mathcal{O}^*$ are countable.

A *data stream* is a countable sequence $\vec{O} = (O_0, O_1, \ldots) \in \mathcal{O}^{\omega}$ of data from $\mathcal{O}$. (Here, ω is the set of natural numbers, so $\mathcal{O}^{\omega}$ is the set of all maps assigning an observable property to every natural number.) We use the following notation: $\vec{O}_n$ is the n-th element in $\vec{O}$; $\vec{O}[n]$ is the initial segment of $\vec{O}$ of length n, $(O_0, \ldots, O_{n-1})$; $set(\vec{O}) := \{O \mid O \text{ is an element of } \vec{O}\}$ is the set of all data in $\vec{O}$; $*$ is the concatenation operator on strings.

The intuition is that at stage n of a data stream, the agent observes the information in O_n. A data stream captures a possible future history of observations in its entirety, while a data sequence captures only a finite part of such a history.

Given a state $s \in S$, a *data stream for s* is a stream $\vec{O} \in \mathcal{O}^{\omega}$ such that $\mathcal{O}_s = \{O \in \mathcal{O} \mid \bigcap_{i=0}^{n} O_i \subseteq O \text{ for some } n \in \omega\}$. Such a stream is "sound" (every data in $\vec{O}$ is true at s) and "complete" (every true data is entailed by some finite set of observations in $\vec{O}$).

Example 1. *Let our epistemic space $\mathbb{S} = (S, \mathcal{O})$ be the real numbers, with observable properties given by open intervals with rational endpoints: $S := \mathbb{R}$, $\mathcal{O} := \{(a, b) \mid a, b \in Q, a \leq b\}$, where $(a, b) := \{x \in \mathbb{R} \mid a < x < b\}$. For instance, observables may represent measurements of a physical quantity (such as a position along a one-dimensional line) that takes real numbers as its possible values. In such case, for any state $x \in \mathbb{R}$ and any two sequences $a_n, b_n \in Q$ of rational numbers, such that $a_n \leq x \leq b_n$ and both sequences converge to x, the sequence $(a_0, b_0), \ldots, (a_n, b_n), \ldots$ is a (sound and complete) data stream for x.*

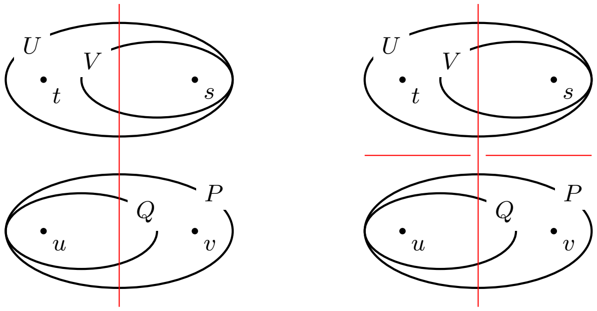


Figure 1: A problem $\mathbb{P}$ (left-hand side) and its refinement $\mathbb{P}'$ (right-hand side), see Example 3

Other examples include standard n-dimensional Euclidean spaces*, e.g., $S = R^3$ with $\mathcal{O}$ consisting of all open balls with rational radius and center.*

Definition 2. *An* inductive problem *is a pair $\mathbb{P} = (\mathbb{S}, \mathcal{Q})$ consisting of an epistemic space $\mathbb{S} = (S, \mathcal{O})$ together with a "question" $\mathcal{Q}$, i.e., a partition[3] of S. The cells A_i of the partition $\mathcal{Q}$ are called* answers*. Given $s \in S$, the unique $A \in \mathcal{Q}$ with $s \in A$ is called* the answer to $\mathcal{Q}$ at s*, and denoted A_s. We say that a problem $\mathbb{P}' = (\mathbb{S}, \mathcal{Q}')$ is a* refinement *of another problem $\mathbb{P} = (\mathbb{S}, \mathcal{Q})$ (or that the corresponding question $\mathcal{Q}'$ is a* refinement *of the question $\mathcal{Q}$) if every answer of $\mathcal{Q}$ is a disjoint union of answers of $\mathcal{Q}'$.*

The *most refined* question concerns the identity of the real world.

Example 2. *The* learning question *on a space S is $\mathcal{Q} = \{\{s\} \mid s \in S\}$ ('What is the actual state?').*

Example 3. *Let $\mathbb{S} = (S, \mathcal{O})$, where $S = \{s, t, u, v)$, $\mathcal{O} = \{U, V, P, Q\}$, with $U = \{s, t\}$, $V = \{s\}$,$P = \{u, v\}$,$Q = \{u\}$. Take the problem $\mathbb{P} = (\mathbb{S}, \mathcal{Q})$, given by the question $\mathcal{Q} = \{\{t, u\}, \{s, v\}\}$ depicted on the left-hand side of Figure 1. This can obviously refined to obtained the problem $\mathbb{P}' = (\mathbb{S}, \mathcal{Q}')$ given by the learning question $\mathcal{Q} = \{\{s\}, \{t\}, \{u\}, \{v\}\}$ for this space, as depicted on the right-hand side of Figure 1.*

3. LEARNING AND PROBLEM-SOLVING

Definition 3. *Let $\mathbb{S} = (S, \mathcal{O})$ be an epistemic space and let $\sigma_0, \ldots, \sigma_n \in \mathcal{O}$. An* agent *(also known as a "learner", or a "learning method") is a map $\mathcal{L}$ that associates to any epistemic space $\mathbb{S}$ and any data sequence $(\sigma_0, \ldots, \sigma_n)$ some family $\mathcal{L}_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n) \subseteq \mathcal{P}(S)$ of subsets of S, satisfying a "consistency" condition: $\emptyset \notin \mathcal{L}_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n)$ whenever $\bigcap_{i=o}^{n} \sigma_i \neq \emptyset$.*

Intuitively, after observing the data sequence $\vec{\sigma} = (\sigma_0, \ldots, \sigma_n)$, we can say that agent $\mathcal{L}$ *believes a proposition P after observing the data sequence $\vec{\sigma} = (\sigma_0, \ldots, \sigma_n)$*, and write $B_{\mathcal{L}}^{\vec{\sigma}} P$ iff $P \in \mathcal{L}_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n)$. We can also interpret this as a *conditional belief*, rather than as revised belief, the agent believes every $P \in \mathcal{L}_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n)$ conditional on $\sigma_0, \ldots, \sigma_n$.

[2] And in the same time (if we adopt a "simplicity" interpretation of the prior), this last result can be seen as a vindication of Ockham's razor (in line with Kevin Kelly's program).

[3] This means that $\bigcup_{i \in I} A_i = S$, and $A_i \cap A_j = \emptyset$ for all $i \neq j$.

But in the end we are of course interested in the actual revised beliefs after observing the data, so the assumption in this case is that conditional beliefs guide the agent's revision strategy: they "pre-encode" future belief revisions, to use a term coined by J. van Benthem [6]. The above consistency simply means that each of the agent's beliefs is consistent whenever the observed data are consistent.

A *doxastic agent* is one whose set $\mathcal{L}_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n)$ of beliefs forms a *(proper) filter* on S when observing consistent data; in other words, her beliefs are (consistent when possible, and also) *inference-closed* (i.e., if $P \subseteq Q$ and $P \in \mathcal{L}_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n)$, then $Q \in \mathcal{L}_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n)$) and *conjunctive* (i.e., if $P, Q \in \mathcal{L}_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n)$ then $(P \cap Q) \in \mathcal{L}_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n)$). Hence, for any doxastic agent $\mathcal{L}$ and every consistent data sequence $\vec{\sigma}$, the belief operator $B^{\vec{\sigma}}_{\mathcal{L}}$ (as defined above) satisfy the usual $KD45$ axioms of doxastic logic.

A *standard agent* is a doxastic agent $\mathcal{L}$ whose beliefs form a *principal filter*, i.e., all her beliefs are entailed by one "strongest belief"; formally, a doxastic agent $\mathcal{L}$ is standard iff for every data sequence $\vec{\sigma}$ over any epistemic space $\mathbb{S}$ there exists some set $L_{\mathbb{S}}(\vec{\sigma})$, such that

$$\mathcal{L}_{\mathbb{S}}(\vec{\sigma}) = \{P \subseteq S \mid L_{\mathbb{S}}(\vec{\sigma}) \subseteq P\}.$$

It is easy to see that in this case, we must have $L_{\mathbb{S}}(\vec{\sigma}) = \bigcap \mathcal{L}_{\mathbb{S}}(\vec{\sigma})$. Indeed, we can equivalently define a doxastic agent $\mathcal{L}$ to be standard iff $\bigcap \mathcal{L}_{\mathbb{S}}(\vec{\sigma}) \in \mathcal{L}_{\mathbb{S}}(\vec{\sigma})$ holds for all data sequences $\vec{\sigma}$. *Standard agents are globally consistent* whenever possible: $\bigcap \mathcal{L}_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n) \neq \emptyset$ whenever $\bigcap_{i=o}^{n} \sigma_i \neq \emptyset$.

Traditional learning methods in Formal Learning Theory correspond to our standard agents, and they are typically identified with the map L (given by $L_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n) := \bigcap \mathcal{L}_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n)$). From now on we follow this tradition, and refer to standard agents using the map L. But in general we do *not* restrict ourselves to standard agents.

An *AGM agent* is an agent $\mathcal{L}^{\leq}$ who forms beliefs by *AGM conditioning*, i.e., it comes endowed with a map that associates any epistemic space $\mathbb{S}$ some total preorder[4] $\leq_{\mathbb{S}}$ on S, called *"prior" plausibility relation*; and whose beliefs after observing any data sequence $\vec{\sigma} = (\sigma_0, \ldots, \sigma_n)$ are given by

$$\mathcal{L}^{\leq}_{\mathbb{S}}(\vec{\sigma}) := \{P \subseteq S \mid \exists s \in \bigcap_{i=0}^{n} \sigma_i \ \forall t \in \bigcap_{i=0}^{n} \sigma_i \ (t \leq s \Rightarrow t \in P)\}.$$

Intuitively, $t \leq s$ means that t is *at least as plausible* as s (according to our agent). So, an AGM agent believes P conditional on a data sequence $\vec{\sigma}$ iff P is true in all the states (consistent with the data) that are "plausible enough".

It is easy to see that *every AGM agent is a doxastic agent*: $\mathcal{L}^{\leq}_{\mathbb{S}}(\vec{\sigma})$ is a proper filter whenever $\bigcap_{i=0}^{n} \sigma_i \neq \emptyset$; hence, the beliefs of an AGM agent satisfy the usual $KD45$ axioms of doxastic logic (when learning any consistent data sequence).

Moreover, it is well-known that in fact, *the beliefs of AGM agents satisfy all the so-called AGM axioms from Belief Revision Theory* [1]: if, for any data sequence $\vec{\sigma} = (\sigma_0, \ldots, \sigma_n)$, we set $T = \mathcal{L}(\sigma_0, \ldots, \sigma_n)$, and for any new observation $\phi \in \mathcal{O}$ we set $T * \phi = \mathcal{L}(\sigma_0, \ldots, \sigma_n, \phi)$, then the resulting revision operator $*$ satisfies all the AGM postulates. In fact, for any AGM agent $\mathcal{L}$, if we interpret the operator $B^{\vec{\sigma}}_{\mathcal{L}}$ (as defined above) as representing a conditional belief $B^{\sigma_0 \wedge \cdots \wedge \sigma_n}$, then the sound and complete logic of these conditional belief operators is the so-called Conditional Doxastic Logic [8, 5] (which is itself just a repackaging of the AGM postulates in the language of conditional logic).

Observation 1. *Given a total preorder $\leq$ on S and a subset $A \subseteq S$, set*

$$Min_{\leq}(A) := \{s \in A \mid s \leq t \ \textit{for all}\ t \in A\}$$

for the set of $\leq$-minimal states in A. Let $\vec{\sigma} = (\sigma_0, \ldots, \sigma_n)$ be any data sequence such that $Min_{\leq}(\bigcap_{i=0}^{n} \sigma_i) \neq \emptyset$. Then $\mathcal{L}^{\leq}_{\mathbb{S}}(\vec{\sigma})$ is the principal filter generated by $Min_{\leq}(\bigcap_{i=0}^{n} \sigma_i)$, i.e., we have

$$\mathcal{L}^{\leq}_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n) := \{P \subseteq S \mid Min_{\leq}(\bigcap_{i=0}^{n} \sigma_i) \subseteq P\}.$$

In general though, the filter $\mathcal{L}^{\leq}_{\mathbb{S}}(\vec{\sigma})$ is not principal. So AGM agents are *not* necessarily standard agents. But there is an important case when they are standard: whenever the preorder $\leq_{\mathbb{S}}$ is well-founded in every space $\mathbb{S}$ (i.e., there are no infinite chains $s_0 > s_1 > s_2 \ldots$ of more and more plausible states). It is easy to see that the map L associated to a standard AGM agent is given by the set of $\leq$-minimal states consistent with the data:

$$L^{\leq}_{\mathbb{S}}(\sigma_0, \ldots, \sigma_n) := Min_{\leq}(\bigcap_{i=0}^{n} \sigma_i).$$

Intuitively, this means that a standard AGM agent believes a proposition P iff P is true in all the "most plausible" states consistent with the data.

The original semantics of AGM belief was given using only standard AGM agents. But this semantics was in fact borrowed by Grove [18] from Lewis' semantics for conditionals [27], which did *not* assume well-foundedness.[5]

Definition 4. *Let $\mathbb{S}$ be an epistemic space. An agent $\mathcal{L}$* verifies a proposition $A \subseteq S$ in the limit *if, for every state $s \in S$ and every data stream $\vec{O}$ for s, we have $s \in A$ iff there exists some $k \in \omega$ such that $A \in \mathcal{L}_{\mathbb{S}}(\vec{O}[n])$ for all $n \geq k$. For standard agents, this means that $L_{\mathbb{S}}(\vec{O}[n]) \subseteq A$ for all $n \geq k$. A set $A \subseteq S$ is* verifiable in the limit *if there exists some agent that verifies A in the limit.*[6]

An agent $\mathcal{L}$ falsifies a proposition $A \subseteq S$ in the limit *if, for every state $s \in S$ and every data stream for $\vec{O}$ for s, we have $s \notin A$ iff there exists some $k \in \omega$ such that $A^c \in \mathcal{L}(\mathbb{S}, \vec{O}[n]) \subseteq A^c$ for all $n \geq k$. (Here, as in the rest of this paper, $X^c := S \setminus X$ stands for the complement of X.) For a standard agent, this means $L(\mathbb{S}, \vec{O}[n]) \subseteq A^c$ for all $n \geq k$,*

A proposition $A \subseteq S$ is falsifiable in the limit *if there exists some agent that falsifies A in the limit.*

A proposition $A \subseteq S$ is decidable in the limit *if it is both verifiable and falsifiable in the limit.*

An agent $\mathcal{L}$ solves a problem $\mathbb{P} = (\mathbb{S}, \mathcal{Q})$ *if, for every state $s \in S$ and every data stream $\vec{O}$ for s, there exists some $k \in \omega$ such that $A_s \in \mathcal{L}_{\mathbb{S}}(\vec{O}[n])$ for all $n \geq k$. (Recall that A_s is true answer to $\mathcal{Q}$ at s.) For a standard agent, this means that $L_{\mathbb{S}}(\vec{O}[n]) \subseteq A_s$ for all $n \geq k$. A problem is* solvable (in the limit) *if there exists some agent that solves it.*

[4] A total preorder on S is a binary relation $\leq$ on S that is reflexive, transitive, and connected (i.e., for all $s, t \in S$, we have either $s \leq t$ or $t \leq s$).

[5] Indeed, Lewis' definition of conditionals has a similar shape to our above definition of (conditional) beliefs for non-standard AGM agents.

[6] For a discussion of the relationship between verifiability and learnability see, e.g., [19, 12].

An epistemic space $\mathbb{S} = (S, \mathcal{O})$ *is* learnable (by an agent $\mathcal{L}$) *if the (problem given by the) learning question* $\mathcal{Q}_S = \{\{s\} \mid s \in S\}$ *is solvable (by* $\mathcal{L}$*).*

All the above notions have a standard *counterpart, e.g., A is* standardly verifiable *if there exist some standard agent that verifies it;* $\mathbb{P}$ *is* standardly solvable *if it can be solved by some standard agent, etc.*

Note that standard learnability is essentially the same as Gold's *identifiability in the limit* [30, 17].

Examples and Counterexamples: An example of *non-learnable* space $\mathbb{S} = (S, \mathcal{O})$ is obtained by taking four abstract states $S = \{s, t, u, w\}$ and two observable properties $\mathcal{O} = \{V, U\}$, with $V = \{s, t, u\}$ and $U = \{t, u, w\}$, as depicted in Figure 2. Since states s and t satisfy the same observable properties, no learning method will ever distinguish them.

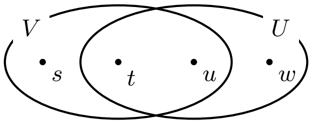


Figure 2: A non-learnable space

But even spaces in which no two states satisfy the same observations can still be non-learnable, e.g., *all the n-dimensional Euclidean spaces from Example 1 are not learnable* (though, as we will see, many questions are solvable and many subsets are decidable over these spaces). Another example of *non-learnable* space is given in Figure 3: formally, $\mathbb{S} = (S, \mathcal{O})$, where $S := \{s_n \mid n \in \omega\} \cup \{s_\infty\}$, and $\mathcal{O} = \{O_i \mid i \in \omega\}$, and for any $i \in \omega$, $O_i := \{s_i, s_{i+1}, \dots\} \cup \{s_\infty\}$.

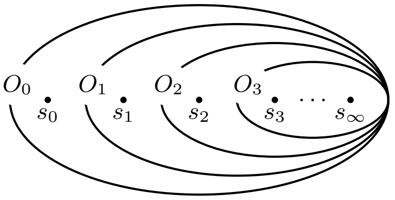


Figure 3: Another non-learnable space

In contrast, an example of *learnable* space is in Figure 4: formally, $S = \{s_n \mid n \in \omega\}$ consists of countably many distinct states, with $\mathcal{O} = \{O_n \mid n \in \omega\}$, where $O_n = \{s_0, s_1, s_2, \dots, s_n\}$. A standard agent that can learn this space in the limit is given by setting $L(\sigma_1, \dots, \sigma_n)$ to be the *maximum number (in the natural order)* in $\bigcap_{i=0}^{n} \sigma_i$, whenever there is such a maximum number, and setting$L(\sigma_1, \dots, \sigma_n) := \bigcap_{i=0}^{n} \sigma_i$ otherwise.

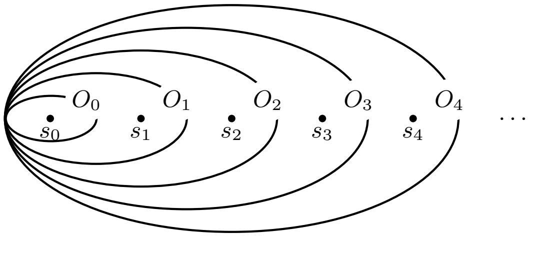


Figure 4: A learnable space

Proposition 1. *Let* $\mathbb{S}$ *be an epistemic space,* $A \subseteq S$ *a proposition and* $\mathbb{P} = (\mathbb{S}, \mathcal{Q})$ *an inductive problem. Then we have the following:*

- *A is verifiable (falsifiable, decidable) in the limit iff it is standardly verifiable (falsifiable, decidable) in the limit.*
- $\mathbb{P}$ *is solvable iff it is standardly solvable.*
- $\mathbb{S}$ *is learnable iff it is standardly learnable.*

Proof. Let $A \subseteq S$ be a set that is verifiable (falsifiable, decidable) by an agent $\mathcal{L}$ on an epistemic space $\mathbb{S}$. We construct a standard agent that does the same thing, by setting, for every data sequence $\vec{\sigma} \in \mathcal{O}^*$: $L_{\mathbb{S}}(\vec{\sigma}) := A$ if $A \in \mathcal{L}_{\mathbb{S}}(\vec{\sigma})$, $L_{\mathbb{S}}(\vec{\sigma}) := A^c$ if $A \notin \mathcal{L}_{\mathbb{S}}(\vec{\sigma})$ but $A^c \in \mathcal{L}_{\mathbb{S}}(\vec{\sigma})$, and $L_{\mathbb{S}}(\vec{\sigma}) := S$ otherwise. Also, on any *other* space $\mathbb{S}' = (S', \mathcal{O}')$, we set by default $L_{\mathbb{S}'}(\vec{\sigma}') := S'$.

Similarly, let $\mathbb{P} = (\mathbb{S}, \mathcal{Q})$ be a problem that is solvable by $\mathcal{L}$. Let $\leq$ be some arbitrary well-order of the set $\mathcal{Q}$. (Such a well-order exists, by the Well-Ordering Theorem.) We construct a standard agent who also solves $\mathbb{P}$, by setting $L_{\mathbb{S}}(\vec{\sigma}) := A$ if A is the first answer in $\mathcal{Q}$ (according to $\leq$) such that $A \in \mathcal{L}_{\mathbb{S}}(\vec{\sigma})$ holds; and $L_{\mathbb{S}}(\vec{\sigma}) := S$ if no such answer exists. (As before, we can extend our agent to any other space $\mathbb{S}' = (S', \mathcal{O}')$, by setting $L_{\mathbb{S}'}(\vec{\sigma}') := S'$.)

By applying this to the learning problem $\mathcal{Q} = \{\{s\} \mid s \in S\}$, we obtain the similar result for learnability. □

In conclusion, everything that can be learned by any agent can also be learned by some standard agent. However, this is no longer true when we restrict to more canonical types of agents (such as AGM agents).

Proposition 2. *There exist spaces that are learnable, but not learnable by standard AGM agents. (Hence, there exist solvable problems that are not solvable by standard AGM agents.)*

Proof. Here is a counterexample from [13, 3, 4]. Take the epistemic model from Figure 4. This space is learnable, and thus learnable by *AGM* conditioning, but it is not learnable by standard conditioning. Indeed, this space is learnable by conditioning *only* with respect to the following *non-wellfounded* prior: $s_0 > s_1 > \dots > s_n > s_{n+1} > \dots$ □

4. THE OBSERVATIONAL TOPOLOGY

In this section, we assume familiarity with the following notions: *topology* τ (identified with its family of *open* subsets) over a set S of points, *topological space* (S, τ), *open* sets, *closed* sets, *interior* $Int(X)$ and *closure* $\overline{X}$ of a set X, (open) *neighborhood* of a point s, *base* of a topology and *local base (of neighborhoods) at a point.* We use letters U, U', etc., for open sets in τ, and letters C, C', etc., for closed sets.

A space is said to be *second-countable* if its topology has a countable base. Given a topological space (S, τ), the *specialization preorder* $\sqsubseteq \subseteq S \times S$ is defined in the following way: for any $s, t \in S$, we set

$$s \sqsubseteq t \quad \text{iff} \quad \forall U \in \tau\, (s \in U \Rightarrow t \in U).$$

Separation Principles. In this paper we use four key topological separation notions. The first is the well-known separation axiom $T0$, which will be satisfied by all the topologies that arise in our setting. The second is the separation axiom TD. This condition (together with countability) will be shown to characterize *learnable* spaces. The next two notions are analogues of TD separation for *questions*. Instead of asking for open sets that separate points (states), these conditions require the existence of open sets that separate *answers* (to the same question). The concept of *locally closed questions* is a first analogue of TD, and it will be shown to characterize in some sense *solvable* problems. Finally, the notion of *linearly separated questions* is a stronger analogue of TD for questions, which characterizes a stronger type of solvability, what we will call *direct solvability by (AGM) conditioning.*

Definition 5. *A topological space (S, τ) satisfies* the separation axiom $T0$ *if the specialization preorder is actually a partial order, i.e., it is antisymmetric: $s \sqsubseteq t \sqsubseteq s$ implies $s = t$. Equivalently, if $s \neq t$, then there exists some "separating" open U, such that either $s \in U$, $t \notin U$, or $s \notin U$, $t \in U$.*

The space (S, τ) satisfies the separation axiom TD *iff for every point $s \in S$, there is an open $U_x \ni x$ such that $y \not\sqsubseteq x$ for all $y \in O_x \setminus \{x\}$. Equivalently: for every $s \in S$ there is an open $U \in \tau$ such that $\{s\} = U \cap \overline{\{s\}}$.*

Essentially, $T0$ says that every two points $s \neq t$ can be separated (by an open U) one way or another (i.e., either $s \in U$, $t \notin U$, or $s \notin U$, $t \in U$), while TD essentially says that every point s can be separated (by an open neighborhood) from all the points $t \neq s$ that are inseparable from s.[7]

Definition 6. *Given a topological space (S, τ), a set $A \subseteq S$ is* locally closed *if it is the intersection $A = U \cap C$ of an open set U with a closed set C. Equivalently, if it is of the form $A = U \cap \overline{A}$ for some open U.*

A set is ω-constructible *if it is a countable union of locally closed sets.*

A question $\mathcal{Q}$ (partition of S) is locally closed *if all its answers are locally closed. A problem $\mathbb{P}$ is locally closed if its associated question is locally closed.*

Essentially, locally closed questions are partitions with the property that every "answer" (i.e., partition cell) A can be separated (by an open neighborhood) from all the non-A-states that are inseparable from A.[8]

Definition 7. *A question $\mathcal{Q}$ is* linearly separated *if there exists some* total order $\trianglelefteq$ *on the answers in $\mathcal{Q}$, such that $A \cap \overline{\bigcup_{B \lhd A} B} = \emptyset$. In other words,* every answer A can be separated (by some open $U_A \supseteq A$) from the union of all the previous answers*: $U_A \cap B = \emptyset$ for all $B \lhd A$.*

Essentially, a linearly separated question is one whose answers can be totally ordered by a "plausibility" (or "simplicity") order, in such a way that every answer A can be separated (by an open neighborhood $U_A \supseteq A$) from all answers that are "more plausible" (or "simpler") than A.

Definition 8. *The* observational topology $\tau_{\mathbb{S}}$ *associated with an epistemic space $\mathbb{S} = (S, \mathcal{O})$ is the topology generated by $\mathcal{O}$ (i.e., the smallest collection of subsets of S, that includes $\mathcal{O} \cup \{\emptyset, S\}$ and is closed under finite intersections and arbitrary unions).*

From now on, we will always implicitly consider our epistemic spaces $\mathbb{S}$ to also be topological spaces $(S, \tau_{\mathbb{S}})$, endowed with their observational topology $\tau_{\mathbb{S}}$. Every topological property possessed by the associated topological space will thus be also attributed to the epistemic space.

Observation 2. *Every epistemic space is $T0$ and second-countable. A (sound and complete) data stream for s is the same as a local neighborhood base at s.*

Proposition 3. *Every ω-constructible set can be written as a disjoint countable union of locally closed sets.*

Proof. In order to prove this, we first recall some standard topological notions and results: A set is called *constructible* if it is a finite disjoint union of locally closed sets. Obviously, all locally closed sets are constructible. It is known that constructible sets form a Boolean algebra, i.e., the family of constructible sets is closed under complementation, finite unions, and finite intersections.

Suppose $A = \bigcup_{i \in \omega} A_i$, where all A_i are locally closed. Then we can rewrite A as a disjoint union $A = \bigcup_{i \in \omega} B_i$, where we have set $B_i = A_i \setminus (\bigcup_{k<i} A_k) = A_i \cap \bigcap_{k<i} A_k^c$, for every i. Since B_i's are generated from locally closed sets using complementation and finite intersections, they must be constructible. Hence, each B_i can be written as disjoint finite unions of locally closed sets $B_i = \bigcup_{1 \leq j \leq i} B_{ij}$. Hence, we can write $A = \bigcup_{i \in \omega} \bigcup_{1 \leq j \leq i} B_{ij}$ as a disjoint countable union of locally closed sets. □

Definition 9. *A* pseudo-stratification *is a finite or ω-long sequence of locally closed sets $\langle A_i \mid i < \lambda \rangle$ (where $\lambda \in \omega \cup \{\omega\}$), which form a partition of S satisfying the following condition:*

$$\text{if } j < i \text{ then either } A_i \cap \overline{A_j} = \emptyset \text{ or } A_i \subseteq \overline{A_j}.$$

Proposition 4. *Every countable locally closed question can be refined to a pseudo-stratification.*

Proof. Suppose $\Pi = \{A_i \mid i \in \omega\}$ is a countable locally closed question (partition of S). We first show the following:

Claim. There exists a family $\{(\Pi_i, <_i) \mid i \in \omega\}$, satisfying

(1) each Π_i is a finite partition of A_i into locally closed sets;

(2) each $<_i$ is a total order on Π_i;

(3) if $j \leq i$, $E \in \Pi_j$, $B \in \Pi_i$, then either $B \subseteq \overline{E}$ or $B \subseteq \overline{E}^c$;

(4) if $B, E \in \Pi_i$, $E <_i B$, then $B \subseteq \overline{E}^c$.

Proof of Claim: We construct $(\Pi_n, <_n)$ by recursion: for $n = 0$, set $\Pi_0 := \{A_0\}$, with $<_0$ trivial. For the step $n+1$: assume given $\{(\Pi_i, <_i) \mid i \leq n\}$ satisfying the above four conditions (for $i \leq n$). We set

$$\Pi_{n+1} := \{B_f \mid f : \bigcup_{i=1}^{n} \Pi_i \to \{0, 1\}\},$$

[7] A point y is "inseparable" from x if every open neighborhood of y contains x, i.e. y and x are in the topological refinement order $y \sqsubseteq x$.

[8] Here, a state t is said to be "inseparable" from a set A if there is no open neighborhood $U \ni t$ that is disjoint from A.

where for each function $f : \bigcup_{i=1}^{n} \Pi_i \to \{0,1\}$ we have set

$$B_f := A_{n+1} \cap \bigcap\{\overline{E} \mid E \in f^{-1}(0)\} \cap \bigcap\{\overline{E}^c \mid E \in f^{-1}(1)\}.$$

It is obvious that the B_f's are locally closed (given that A_{n+1} is locally closed) and that they form a partition of A_{n+1}. So condition (1) is satisfied.

It is also easy to check condition (2) for $i = n+1$: let $j < n+1$, $E \in \Pi_j$ and $B_f \in \Pi_{n+1}$. Then we have either $f(E) = 0$, in which case $B_f \subseteq \overline{E}$ (by construction of B_f), or else $f(E) = 1$, in which case $B_f \subseteq \overline{E}^c$.

To construct the order $<_{n+1}$, observe first that there is a natural total order $<^{(n)}$ on the disjoint union $\bigcup_{i=1}^{n} \Pi_i$, namely the one obtained by concatenating the orders $<_0$, $<_1$, ... , $<_n$. (More precisely, if, for every $B \in \bigcup_{i=1}^{n} \Pi_i$, we set $i(B)$ to be the unique index $i \leq n$ such that $B \in \Pi_i$, then the order $<^{(n)}$ is given by setting: $B <^{(n)} E$ iff either $i(B) < i(E)$, or else $i(B) = i(E)$ and $B <_{i(B)} E$.)

Now, the order $<_{n+1}$ on B_f's is given by the lexicographic order induced by $<^{(n)}$ on the functions f (thought as "words" written with the letters 0 and 1). More precisely, we set:

$$B_f <_{n+1} B_g$$

iff there exists some set $E \in \bigcup_{i=1}^{n} \Pi_i$ such that

$$\Big(\forall E' <^{(n)} E \, f(E') = g(E'), \text{ but } f(E) < g(E)\Big),$$

where $<$ is the usual order $0 < 1$ on $\{0,1\}$. Clearly, $<_{n+1}$ is a total order on Π_{n+1}, so condition (2) is satisfied.

Finally, we check condition (4) for $n+1$, let $B_f, B_g \in \Pi_{n+1}$ such that $B_f <_{n+1} B_g$. By definition of the order $<_{n+1}$, this means that there exists some $E \in \bigcup_{i=1}^{n} \Pi_i$ such that for all $E' <^{(n)} E$ we have $f(E') = g(E')$ but $f(E) < g(E)$, i.e., $f(E) = 0$ and $g(E) = 1$. By the construction of B_f's, $f(E) = 0$ implies that $B_f \subseteq \overline{E}$, from which we get $\overline{B_f} \subseteq \overline{E}$, and thus $\overline{E}^c \subseteq \overline{B_f}^c$. Similarly, $g(E) = 1$ implies that $B_g \subseteq \overline{E}^c$. So we have $B_g \subseteq \overline{E}^c \subseteq \overline{B_f}^c$, and thus by transitivity of inclusion we get $B_g \subseteq \overline{B_f}^c$. This completes the proof of our Claim.

Given now the above Claim, we can prove our Lemma by taking as our refined partition

$$\Pi' := \bigcup_{i \in \omega} \Pi_i.$$

Clearly, Π' is a refinement of Π consisting of locally closed sets. We now define a well-order $<'$ on Π' as the concatenation of all the $\leq_i$'s.[9] Obviously, $<'$ is a total order of type $\leq \omega$ on Π', so we get finite or ω-long sequence that enumerates Π'. The above properties (3) and (4) ensure that this is a pseudo-stratification. □

Lemma 1. *Given a pseudo-stratification $\langle A_i \mid i < \lambda \rangle$ (of length $\lambda \leq \omega$), there exists a λ-long sequence of open sets $\langle U_i \mid i < \lambda \rangle$, satisfying:*

(1) $U_i \cap \overline{A_i} = A_i$;

(2) if $j < i$ and $U_i \cap A_j \neq \emptyset$, then $A_i \subseteq \overline{A_j}$.

Proof. We know that each A_i is locally closed, so there exists some open set $U^{A_i} \in \tau$ such that $U^{A_i} \cap \overline{A_i} = A_i$. Now, for all $i \in \omega$ set

$$U_i := U^{A_i} \cap \bigcap\{\overline{A_j}^c \mid j < i, A_i \subseteq \overline{A_j}^c\}.$$

Let us first check that the sequence $\langle U_i \mid i < \lambda \rangle$ satisfies condition (1):

$$U_i \cap \overline{A_i} = (U^{A_i} \cap \bigcap\{\overline{A_j}^c \mid j < i, A_i \subseteq \overline{A_j}^c\}) \cap \overline{A_i}$$

$$= (U_i \cap \overline{A_i}) \cap \bigcap\{\overline{A_j}^c \mid j < i, A_i \subseteq \overline{A_j}^c\}$$

$$= A_i \cap \bigcap\{\overline{A_j}^c \mid j < i, A_i \subseteq \overline{A_j}^c\} = A_i$$

Second, let us check condition (2): Suppose that we have $j < i$ and $U_i \cap A_j \neq \emptyset$, but $A_i \not\subseteq \overline{A_j}$. Since $(A_i)_{i<\lambda}$ is a pseudo-stratified sequence, from $j < i$ and $A_i \not\subseteq \overline{A_j}$ we can derive $A_i \subseteq \overline{A_j}^c$. By the construction of U_i, this implies that $U_i \subseteq \overline{A_j}^c$, and hence that $U_i \cap A_j \subseteq \overline{A_j}^c \cap A_j \subseteq \overline{A_j}^c \cap \overline{A_j} = \emptyset$, which contradicts the assumption that $U_i \cap A_j \neq \emptyset$. □

Lemma 2. *Every pseudo-stratification is linearly separated.*

Proof. Let $\Pi = \{A_i \mid i < \lambda\}$ be a pseudo-stratification (with $\lambda \leq \omega$), and let $\langle U_i \mid i < \lambda \rangle$ be a sequence satisfying the conditions of Lemma 1. It is clear that, in order to prove our intended result, it is enough to construct a total order $\trianglelefteq$ on the set $\{i \in \omega | i < \lambda\} = \lambda \subseteq \omega$, such that

$$U_i \cap A_j \neq \emptyset \Rightarrow i \trianglelefteq j.$$

For this, we first define a reflexive relation R on λ, by setting

$$iRj \iff U_i \cap A_j \neq \emptyset.$$

Claim: There are no non-trivial cycles

$$i_1 R \cdots i_n R i_1 \text{ (with distinct } i_k\text{'s).}$$

Proof of Claim: Let $i_1 R \cdots i_n R i_1$ be a non-trivial cycle of *minimal length* $n \geq 2$. There are two cases:

Case 1: $n = 2$, i.e., $i_1 R i_2 R i_1$ with $i_2 \neq i_1$. We must have either $i_1 < i_2$ or $i_2 < i_1$. Without loss of generality, we can assume $i_1 < i_2$ (otherwise, just swap i_1 and i_2, and use the cycle $i_2 R i_1 R i_2$). From $i_2 R i_1$, we get $U_{i_2} \cap A_{i_1} \neq \emptyset$. This together with $i_1 < i_2$, gives us $A_{i_2} \subseteq \overline{A_{i_1}}$ (by condition (2) from Lemma 2), and hence $U_{i_1} \cap A_{i_2} \subseteq U_{i_1} \cap \overline{A_{i_1}} = A_{i_1}$. From this, we get that $U_{i_1} \cap A_{i_2} = (U_{i_1} \cap A_{i_2}) \cap A_{i_2} \subseteq A_{i_1} \cap A_{i_2} = \emptyset$ (since $i_1 \neq i_2$, so A_{i_1} and A_{i_2} are different answers, hence disjoint), so we conclude that $U_{i_1} \cap A_{i_2} = \emptyset$. But on the other hand, from $i_1 R i_2$ we get $U_{i_1} \cap A_{i_2} \neq \emptyset$. Contradiction.

Case 2: $n > 2$. Since all the i_k's are distinct, there must exist a (unique) smallest index in the cycle. Without loss of generality (since otherwise we can rearrange the indices, permuting the cycle), we can assume that i_3 is the smallest index. (Note that, since $n > 2$, there must be at least three distinct successive indices i_1, i_2, i_3.) So $i_3 < i_1$ and $i_3 < i_2$. From $i_2 R i_3$ we get $U_{i_2} \cap A_{i_2} \neq \emptyset$. Since $i_3 < i_2$, it follows that $A_{i_2} \subseteq \overline{A_{i_3}}$ (by Lemma 2). But on the other hand, $i_1 R i_2$ gives us $U_{i_1} \cap A_{i_2} \neq \emptyset$. We hence obtain $U_{i_1} \cap \overline{A_{i_3}} \neq \emptyset$. This, together with $i_3 < i_1$, gives us $A_{i_1} \subseteq \overline{A_{i_3}}$ (again by Lemma 2). From this, we derive $A_{i_1} \subseteq U_{i_1} \cap \overline{A_{i_3}}$ (since $A_i \subseteq U_i$ for

[9] Once again, one can specify this more precisely by first defining $i : \Pi' \to \omega$ by choosing $i(B)$ to be the unique index i such that $B \in \Pi_i$, and finally defining: $B <' E$ iff either $i(B) < i(E)$, or else $i(B) = i(E)$ and $B <_{i(B)} E$.

all i). Let now $s \in A_{i_1}$ be any state satisfying the answer $A_{i_1} \subseteq U_{i_1} \cap \overline{A_{i_3}}$. So we have $s \in U_{i_1}$ and $s \in \overline{A_{i_3}}$, which together imply that $U_{i_1} \cap A_{i_3} \neq \emptyset$ (since $s \in \overline{A_{i_3}}$ implies that every open neighborhood of s intersects A_{i_3}). Hence, we have $i_1 R i_3$, which means we can shorten the cycle by eliminating i_2, we obtain contradiction.

Given the above Claim, it follows that the transitive closure R^* is a partial order on λ (which obviously includes R). By the Order Extension Principle, we can extend R^* to a total order $\trianglelefteq$ on λ, which still includes R. □

5. TOPOLOGICAL CHARACTERIZATION OF SOLVABILITY

Definition 10. *Let $\mathbb{S} = (S, \mathcal{O})$ be an epistemic space, L be a standard agent, $A \subseteq S$, and $s \in A$. An* A-locking sequence for s (with respect to L) *is a data sequence $\sigma = (O_1, \dots, O_k)$, such that:*

(1) σ is sound for s, i.e., $s \in \bigcap_{1 \leq i \leq k} O_i$;

(2) if δ is any data sequence sound for s, then $L(\mathbb{S}, \sigma * \delta) \subseteq A$.

For a given data sequence σ, we denote by L^{σ}_A the set of all states in A having σ as an A-locking sequence, i.e.,

$$L^{\sigma}_A := \{s \in A \mid \sigma \text{ is an } A\text{-locking sequence for } s \text{ wrt } L\}.$$

Lemma 3. *If A is verifiable in the limit by a standard agent L, then $\bigcup_{\sigma \in \mathcal{O}^*} L^{\sigma}_A = A$.*

Proof. Suppose not. Let A be verifiable in the limit, but such that $A \neq \bigcup_{\sigma \in \mathcal{O}^*} L^{\sigma}_A$. Since all $L^{\sigma}_A \subseteq A$, his means that $A \not\subseteq \bigcup_{\sigma \in \mathcal{O}^*} L^{\sigma}_A$, i.e., there exists some state $s \in A$ for which there is no A-locking sequence. This means that every data sequence σ that is sound for s can be extended to a sequence δ that is also sound for s and has $L(\delta) \not\subseteq A$.

Let now $\vec{O}$ be a (sound and complete) data stream for s. We construct a new infinite data stream $\vec{V}$, by defining increasingly longer initial segments δ_k of $\vec{O}$, in countably many stages: we first set $V_0 = O_0$, thus obtaining an initial segment $\delta_0 = (O_0) = (V_0)$; at the $k+1$-th stage, given some initial segment $\delta_k = (V_0, V_1, \dots, V_{n_k})$ (of some length n_k), we built our next initial segment by taking any extension δ_{k+1} of the sequence $\sigma_k = (V_0, \dots, V_{n_k}, O_{n+1})$ that is sound for s and has $L(\sigma_k) \not\subseteq A$. The resulting infinite stream $\vec{V}$ is a (sound and complete) stream for s (the completeness of $\vec{V}$ with respect to s follows the fact that this stream includes all the elements of $\vec{U}$), but which contains arbitrarily long initial segments σ_k with $L(\sigma_k) \not\subseteq A$. Since $s \in A$, this contradicts the assumption that A is verifiable in the limit. □

Lemma 4. *If $A \subseteq S$ is verifiable in the limit by a standard agent L, then for every data sequence $\sigma = (O_1, \dots O_k)$, the set L^{σ}_A is locally closed.*

Proof. Let $O := \bigcap_{i=1}^{k} O_i$ be the intersection of all the observations in σ. We will show that

$$O \cap \overline{L^{\sigma}_A} = L^{\sigma}_A,$$

from which the desired conclusion follows.

($\supseteq$) If $s \in L^{\sigma}_A$, then σ is an A-locking sequence for s, hence σ is sound for s, and thus $s \in \bigcap_{i=1}^{n} O_i = O$.

($\subseteq$) Suppose that $s \in O \cap \overline{L^{\sigma}_A}$. We prove two claims:

Claim 1: For every data sequence δ that is sound for s and extends σ, we have $L_{\mathbb{S}}(\delta) \subseteq A$.

Proof of Claim 1: Let $\delta = (\delta_1, \dots, \delta_n)$ be a data sequence that is sound for s (i.e., $s \in \delta_i$ for all $i = 1, \dots, n$) and extends σ, i.e., $n \geq k$ and $U_i = O_i$ for all $i \leq k$). Hence, $\bigcap_{i=1}^{n} \delta_i$ is an open neighborhood of s, and $s \in \overline{L^{\sigma}_A}$, so there must exist some $t \in \bigcap_{i=1}^{n} \delta_i$ such that $t \in L^{\sigma}_A$. Hence, $t \in A$ and σ is an A-locking sequence for t. But δ extends σ and is sound for t, so (by the definition of σ being an A-locking sequence for t), we have that $L(\delta) \subseteq A$, which concludes the proof of Claim 1.

Claim 2: We have $s \in A$.

Proof of Claim 2: Let $\vec{V}$ be a stream for s that extends σ (such a stream must exist, since σ is sound for s: just take any stream for s and prefix it with σ). Then, for every $n \geq k$, the sequence $\delta_n = (V_1, \dots, V_n)$ is sound for s and extends σ. Hence, by the above Claim, we must have that $L_{\mathbb{S}}(V_1, \dots, V_n) \subseteq A$ for all $n \geq k$. But we assumed that A is verifiable in the limit, so we must have $s \in A$, which concludes the proof of Claim 2.

From Claims 1 and 2 together, we conclude that σ is an A-locking sequence for $s \in A$, hence $s \in L^{\sigma}_A$. □

Theorem 1. *Given an epistemic space $(S, \mathcal{O})$, a set $A \subseteq S$ is verifiable in the limit iff it is ω-constructible.*

Proof. ($\Leftarrow$) Assume $A = \bigcup_n (U_n \cap C_n)$ is a countable disjoint union of (mutually disjoint) locally closed sets $U_n \cap C_n$ (with U_n open and C_n closed). We define a standard agent L for A on finite data sequences $\delta = (O_1, \dots, O_k)$, by setting $L(\mathbb{S}, \delta) = A^c$, if we have $\bigcap_j O_j \not\subseteq U_n$ for all $n \in \omega$; $L(\mathbb{S}, \delta) = A^c$ (where A^c is the complement of A), if $\bigcap_j O_j \subseteq C^c_n$ holds for the first index $n \in \omega$ such that $\bigcap_j O_j \subseteq U_n$; and $L(\mathbb{S}, \delta) = A$ otherwise. Then it is easy to see that L verifies A in the limit.

($\Rightarrow$) Suppose that A is verifiable in the limit. By Proposition 1, it is then verifiable by a standard agent L. By Lemma 1, A is the union of all sets L^{σ}_A for all finite data sequences σ. But there are only countably many such sequences, so this is a countable union. Moreover, by Lemma 2, each L^{σ}_A is locally closed. Hence A is a countable union of locally closed sets, i.e., an ω-constructible set. □

Corollary 1. *A is decidable in the limit iff both A and A^c are ω-constructible.*

Proof. Follows trivially from the above results. □

Theorem 2. *Let $\mathbb{P} = (\mathbb{S}, \mathcal{Q})$ be an inductive problem on an epistemic space $\mathbb{S}$. The following are equivalent:*

(1) $\mathbb{P}$ is solvable (in the limit);

(2) the associated question $\mathcal{Q}$ is an (at most) countable family of ω-constructible answers;

(3) $\mathcal{Q}$ has an (at most) countable locally closed refinement.

Proof. (1) $\Rightarrow$ (2) : Let $\mathbb{P}$ be a solvable problem. By Proposition 1, there exists some standard agent that solves it. Let L be such a standard agent that solves $\mathbb{P}$.

Claim: Every answer $A \in \mathcal{Q}$ is verifiable in the limit.

Proof of Claim: Let $A \in \mathcal{Q}$ be an answer. We construct a standard agent L^A that verifies it, by setting $L^A_{\mathbb{S}}(\sigma) := A$

iff $L_{\mathbb{S}}(\sigma) \subseteq A$, and $L^A_{\mathbb{S}}(\sigma) := A^c$ otherwise. It is easy to see that L^A verifies A.

Using the Claim and Lemma 3, we obtain that, for each answer $A \in \mathcal{Q}$, there exists some data sequence $\sigma \in \mathcal{O}^*$ such that $L_{\mathbb{S}}(\sigma) \subseteq A$. But $\mathcal{O}^*$ is countable, so there can be only countably many answers in $\mathcal{Q}$.

By the claim above, Lemma 3 and Lemma 4, we obtain that every answer $A \in \mathcal{Q}$ is a countable union of locally closed sets, hence it is ω-constructible.

(2) $\Rightarrow$ (3) : By (2), $\mathcal{Q}$ is (at most) countable, say $\mathcal{Q} = \{A_i \mid i \in \omega\}$, and also each answer $A_I \in \mathcal{Q}$ is ω-constructible, hence it can be written as a countable disjoint union of locally closets $A = \bigcup_{k\in\omega} A_i^k$ (where all A_i^k's locally closed and mutually disjoint). Then the question $\{A_i^k \mid i \in \omega, k \in \omega\}$ is a refinement of $\mathcal{Q}$, which is countable and locally closed.

(3) $\Rightarrow$ (1) : Let $\mathcal{Q}' = \{B_i \mid i \in \omega\}$ be a countable closed refinement of $\mathcal{Q}'$. By Corollary 1, every answer $B \in \mathcal{Q}'$ is decidable, and so by Proposition 1, we can choose for each $B_i \in \mathcal{Q}$ some standard agent L_i that decides B_i. We define now a new standard agent L, by:

$$L_{\mathbb{S}}(\sigma) := \bigcup \{B_i \mid i \in \omega \text{ such that } L_i(\sigma) \subseteq B_i\}.$$

It is easy to see that this agent L solves $\mathcal{Q}'$, and since $\mathcal{Q}'$ is a refinement of $\mathcal{Q}$, L also solves $\mathcal{Q}$. □

Corollary 2. *An epistemic space $\mathbb{S} = (S, \mathcal{O})$ is learnable in the limit iff it is countable and satisfies the TD separation axiom.*

Proof. Apply Theorem 2 to the learning question $\{\{s\} \mid s \in S\}$, noticing that the fact that all its answers are ω-constructible is equivalent to all singletons being locally closed, which is just another formulation of the TD axiom. □

6. UNIVERSALITY OF CONDITIONING

Our aim in this section is to show that AGM conditioning is "universal": every solvable problem can be solved by some AGM agent. First, we introduce an auxiliary notion, that of a problem being *directly solvable by AGM conditioning.*

Given a question $\mathcal{Q}$ on an epistemic space $(S, \mathcal{O})$, any total order $\trianglelefteq \subseteq \mathcal{Q} \times \mathcal{Q}$ on (the answers of) the question $\mathcal{Q}$ induces in a canonical way a total preorder $\leq \subseteq S \times S$, obtained by:

$$s \leq t \ \text{ iff } \ A_s \trianglelefteq A_t$$

(where A_s is the unique answer $A_s \in \mathcal{Q}$ such that $s \in A_s$).

Definition 11. *A problem $\mathbb{P} = (\mathbb{S}, \mathcal{Q})$ is* directly solvable by conditioning *if it is solvable by AGM conditioning with respect to (a prior $\leq$ that is canonically induced, as explained above, by) a total order $\trianglelefteq \subseteq \mathcal{Q} \times \mathcal{Q}$ on (the answers of) the question $\mathcal{Q}$.*

Direct solvability by conditioning essentially means that the problem can be solved by a conditioning agent *who does not attempt to refine the original question*: she forms beliefs only about the answers to the given question, and is thus indifferent between states satisfying the same answer. Direct solvability by conditioning is thus a very stringent condition, and unsurprisingly this form of conditioning is *not* universal.

Proposition 5. *(*K. Genin, personal communication*) Not every solvable problem is directly solvable by conditioning.*

Proof. Let $\mathbb{P}$ be the problem in Example 3, depicted on the left-hand side of Figure 1. It is easy to see that this problem cannot be directly solvable by conditioning! (Indeed, if $\{t, u\} \lhd \{s, v\}$ then v is not learnable by $\lhd$-conditioning; if $\{s, v\} < \{t, u\}$ then t is not learnable by $\lhd$-conditioning; while if $\{t, u\}$ and $\{s, v\}$ are equally plausible, then neither t nor v are learnable.)

But $\mathbb{P}$ *can* be refined to a directly solvable problem, namely the "learning question" $\mathbb{P}'$ (depicted on the right-hand side of Figure 1), which *can* be directly solvable (e.g. if we set $\{t\} \lhd \{s\} \lhd \{v\} \lhd \{u\}$). As a consequence, $\mathbb{P}$ can itself be solved by (non-direct) conditioning (with respect to the order $t < s < v < u$). □

This counterexample suggests a way to prove our intended universality result: it is enough to show that every solvable problem has a refinement that is directly solvable by conditioning. To do this, we first need a structural characterization of direct solvability.

Lemma 5. (Topological Characterization of Direct Solvability by Conditioning) *A problem $\mathbb{P} = (\mathbb{S}, \mathcal{Q})$ is directly solvable by conditioning iff $\mathcal{Q}$ is linearly separated.*

Proof. *Left-to-right implication*: Suppose that $\mathbb{P}$ is directly solvable by conditioning with respect to (a prior $\leq$ that is canonically induced by) a total order $\trianglelefteq \subseteq \mathcal{Q} \times \mathcal{Q}$. Then, for every $s \in S$ choose some sound and complete data stream $\vec{O}^S = (O^s_n)_{n\in\omega}$ for s (with $O^n_s \in \mathcal{O} \subseteq \tau_{\mathbb{S}}$). Direct solvability by conditioning implies then that there exists some N_s such that $Min_{\leq}(O^s_1, \ldots, O^s_{N_s}) \subseteq A_s$. Set $U_s := \bigcap_{i=1}^{N_s} O^s_i \in \tau_{\mathbb{S}}$, so that we have $s \in U_s$ and $Min_{\leq} U_s \subseteq A_s$. Then set $U_A := \bigcup_{s\in A} U_s \in \tau_{\mathbb{S}}$ for every answer $A \in \mathcal{Q}$. We claim that U_A "separates" A from the union of all the answers $B \lhd A$ (as linear separation demands): indeed, by the construction of U_A, it is obvious that (1) $A \subseteq U_A$, and also that $Min_{\leq} U_A \subseteq A$. By unfolding the last clause in terms of $\trianglelefteq$, we obtain that: $A \trianglelefteq B$ holds for all $B \in \mathcal{Q}$ such that $U_A \cap B \neq \emptyset$. Since $\trianglelefteq$ is a total order on $\mathcal{Q}$, this is equivalent to: (2) $U_A \cap B = \emptyset$ for all $B \lhd A$. By (1) and (2) together, we obtain that $\mathcal{Q}$ is linearly separated.

Right-to-left implication: Suppose $\mathcal{Q}$ is linearly separated. Let $\trianglelefteq$ be a total order on $\mathcal{Q}$ that linearly separates it. This means that, for every answer $A \in \mathcal{Q}$, there exists some open set $U_A \in \tau_{\mathbb{S}}$ such that $A \subseteq U_A$ and $U_A \cap B = \emptyset$ for all $B \lhd A$. For each $s \in S$, we set $U_s := U_{A_s}$ (where A_s is the unique answer $A_s \in \mathcal{Q}$ with $s \in A_s$).

Let $\leq$ be the total preorder on S canonically induced by the order $\trianglelefteq \subseteq \mathcal{Q} \times \mathcal{Q}$ (by $s \leq t$ iff $A_s \trianglelefteq A_t$). We show now that $\mathbb{P}$ *is directly solvable by conditioning with respect to* $\leq$. For this, let $s \in S$ be any state, and $\vec{O} = (O_n)_{n\in\omega}$ be a sound and complete stream for s. Completeness of the stream implies that there must exist some $N \in \omega$ such that $\bigcap_{i=1}^{N} O_i \subseteq U_s$.

To conclude our proof, it is enough to show the following

Claim: For every $n \geq N$, we have

$$s \in Min_{\leq}(\bigcap_{i=1}^{n} O_i) \subseteq A_s.$$

First, let us see why this Claim is enough to give us direct solvability by conditioning. The fact that $s \in Min_{\leq}(\bigcap_{i=1}^{n} O_i)$ implies that $Min_{\leq}(\bigcap_{i=1}^{n} O_i) \neq \emptyset$, for all $n \geq N$. A previous observation tells us that, when applied to such data streams, the AGM agent $\mathcal{L}^{\leq}$ produces a "principal filter", given by

$$\mathcal{L}^{\leq}(O_1, \ldots, O_n) = \{P \subseteq S \mid Min_{\leq}(\bigcap_{i=1}^{n} O_i) \subseteq P\}.$$

By the Claim above we have $Min_{\leq}(\bigcap_{i=1}^{n} O_i) \subseteq A_s$, and hence we obtain $A_s \in \mathcal{L}^{\leq}(O_1, \ldots, O_n)$, for all $n \geq N$.

Proof of Claim: Let $n \geq N$. To prove the Claim, it is enough to show the following two implications (for all states t):

(1) $t \in \bigcap_{i=1}^{n} O_i \Rightarrow s \leq t$;

(2) $t \in Min_{\leq}(\bigcap_{i=1}^{n} O_i) \Rightarrow A_t = A_s$.

To show (1), let $t \in \bigcap_{i=1}^{n} O_i$. Then $t \in U_s$ (since $\bigcap_{i=1}^{n} O_i \subseteq \bigcap_{i=1}^{N} O_i \subseteq U_s$), so $U_s \cap A_t \neq \emptyset$. Hence (by linear separation) we must have $A_s \trianglelefteq A_t$, i.e., $s \leq t$.

To show (2), let $t \in Min_{\leq}(\bigcap_{i=1}^{n} O_i)$. This implies that $t \leq s$ (since $s \in \bigcap_{i=1}^{n} O_i$). But by (1), we also have $s \leq t$, and hence $s \leq t \leq s$. This means that $A_s \trianglelefteq A_t \trianglelefteq A_s$. But $\trianglelefteq$ is a total order on $\mathcal{Q}$, so it follows that $A_t = A_s$. □

THEOREM 3. *AGM conditioning is a* universal problem-solving method, *i.e., every solvable problem is solvable by some AGM agent.*

PROOF. Let $\mathbb{P}$ be a solvable problem. From Theorem 2, Proposition 3 and Lemma 2, it follows that $\mathbb{P}$ has a linearly separated refinement $\mathbb{P}'$. By Lemma 5, that refinement is (directly) solvable by an AGM agent $L^{\leq}$. It is obvious (from the definition of solvability) that any doxastic agent which solves the more refined problem $\mathbb{P}'$ solves also the original problem $\mathbb{P}$. □

COROLLARY 3. *AGM conditioning is a* universal learning method, *i.e.,* every learnable space is learnable by some AGM agent.

PROOF. Apply the previous result to the finest question $\mathcal{Q} := \{\{s\} \mid s \in S\}$. □

In contrast, recall that the counterexample in Proposition 2 showed that *standard* AGM agents have a very limited problem-solving power. Standard conditioning is not a universal learning method (while general *AGM* conditioning is universal). This means that *allowing prior plausibility orders that are non-wellfounded is essential for achieving universality of conditioning.* Beliefs generated in this way may occasionally fail to be globally consistent. (Indeed, note that in the counterexample from Proposition 2, the beliefs of the non-standard *AGM* agent who learns the space are initially globally inconsistent. In conclusion, *occasional global inconsistencies are the unavoidable price for the universality of AGM conditioning.*

7. CONCLUSIONS AND CONNECTIONS TO OTHER WORK

The general topological setting for problem-solving assumed here is a variation of the one championed by Kelly in various talks [23] and in unpublished work [24, 25], though until recently we did not realize this close similarity. Our topological characterizations of verifiable, falsifiable and decidable properties are generalizations of results by Kelly [19], who proved characterizations for the special case of Baire spaces.[10] Our result on learning-universality (Corollary 3) is also a generalization of analogue results by Kelly [21, 20], and Kelly, Schulte and Hendricks [26]. But our generalization to arbitrary spaces is highly non-trivial, requiring the use of the TD characterization. (In contrast, the Baire space satisfies the much stronger separation axiom $T1$, which trivializes the specialization order, and so the proof of learning-universality is much easier in this special case: *any* total ω-like ordering of the space can be used for conditioning.) Nevertheless, in a sense, this result is just a topological repackaging of one of our own previous results [13, 3, 4].

While writing this paper, we learned that our TD characterization of learnability (Corollary 2) was independently reproven by Konstantin Genin ([11], unpublished manuscript), soon after we announced its proof. This characterization is actually a topological translation of a classical characterization of identifiability in the limit [2], and in fact it also follows from a result by de Brecht and Yamamoto [9], who prove it for so-called "concept spaces".

Our key new results are far-reaching and highly non-trivial: the topological characterization of solvability (Theorem 2), and the universality of AGM condition for problem-solving (Theorem 3). They required the introduction of new topological concepts (e.g., pseudo-stratifications and linearly separated partitions), and some non-trivial proofs of new topological results.

Philosophically, the importance of these results is that, on the one hand they fully vindicate the general topological program in Inductive Epistemology started by Kelly and others [19, 31], and on the other hand they reassert the power and applicability of the AGM Belief Revision Theory against its critics.

To this conclusion, we need to add an important proviso: our results show that, in order to achieve problem-solving universality, AGM agents need to (a) be "creative", by *going beyond the original problem* (i.e., finding a more refined problem that can be solved directly, and forming prior beliefs about the answer to this more refined question), and (b) admit *non-standard priors*, which occasionally will lead to *beliefs that are globally inconsistent* (although still locally consistent). Such occasional global inconsistencies can give rise to a type of "infinite Lottery Paradox". But this is the price that AGM agents *have to pay* in order to be able to solve every solvable question.

Whether or not this is a price that is worth paying is a different, more vague and more "ideological" question, although a very interesting one. But this question lies beyond the scope of this paper.

8. ACKNOWLEDGMENTS

We thank Johan van Benthem, Nick Bezhanishvili, Konstantin Genin, Thomas Icard and Kevin Kelly for their useful feedback on issues related to this paper. Johan helped us place belief-based learning within the larger context of long-term doxastic protocols [7], and beyond this he gave

[10] In unpublished work [25] the authors claim a characterization of solvability in a general setting. Their characterization is sightly "looser" than ours, and can be easily obtained from ours. Our tighter characterization is the one needed for proving universality.

us his continuous support and encouragement for our work on this line of inquiry. Nick pointed to us the connections between our work and the notions of TD-space and locally closed set. Konstantin pointed to us the connections to the notion of stratification and gave the counterexample proving Proposition 5. His critical feedback on our early drafts was really essential for clarifying our thoughts and cleaning up our proofs, and so it's fair to say that this paper in its current form owes a lot to Konstantin Genin. Thomas Icard's comments on a previous draft and our friendly interactions with him on related topics during our Stanford visits are very much appreciated. Finally, Kevin Kelly's work forms of course the basis and the inspiration for ours. Our frequent discussions with him in recent years influenced the development of our own perspective on the topic. He also gave us excellent reference tips concerning the history of the connections between topology and formal epistemology, as well as concerning his more recent work on related issues.

Nina Gierasimczuk's work on this paper was funded by an Innovational Research Incentives Scheme Veni grant 275-20-043, Netherlands Organisation for Scientific Research (NWO). Sonja Smets was funded in part by an Innovational Research Incentives Scheme Vidi grant from NWO, and by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement no. 283963.

9. REFERENCES

[1] C. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *JSL*, 50:510–530, 1985.

[2] D. Angluin. Inductive inference of formal languages from positive data. *Inform. Control*, 45(2):117–135, 1980.

[3] A. Baltag, N. Gierasimczuk, and S. Smets. Belief revision as a truth-tracking process. In K. Apt, editor, *Proc. of TARK'11*, pages 187–190. ACM, 2011.

[4] A. Baltag, N. Gierasimczuk, and S. Smets. Truth tracking by belief revision. Technical report, ILLC Report PP-2014-20, 2015. To appear in Studia Logica.

[5] A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Proc. of LOFT'7*, number 3 in Texts in Logic and Games, pages 9–58. Amsterdam University Press, 2008.

[6] J. v. Benthem. *Logical Dynamics of Information and Interaction.* Cambridge University Press, 2011.

[7] J. v. Benthem, J. Gerbrandy, T. Hoshi, and E. Pacuit. Merging frameworks for interaction. *JPL*, 38(5):491–526, 2009.

[8] O. Board. Dynamic interactive epistemology. *Games and Economic Behavior*, 49(1):49–80, 2004.

[9] M. de Brecht and A. Yamamoto. Topological properties of concept spaces (full version). *Inform. Comput.*, 208(4):327–340, 2010.

[10] C. Dégremont and N. Gierasimczuk. Finite identification from the viewpoint of epistemic update. *Inform. Comput.*, 209(3):383–396, 2011.

[11] K. Genin. Linearizing a countable TD space. *Unpublished manuscript*, 2015.

[12] N. Gierasimczuk. Identification through inductive verification. In *Proc of TBiLLC'07*, volume 5422 of *LNCS*, pages 193–205. Springer, 2009.

[13] N. Gierasimczuk. *Knowing One's Limits. Logical Analysis of Inductive Inference.* PhD thesis, Universiteit van Amsterdam, The Netherlands, 2010.

[14] N. Gierasimczuk, D. de Jongh, and V. F. Hendricks. Logic and learning. In A. Baltag and S. Smets, editors, *Johan van Benthem on Logical and Informational Dynamics.* Springer, 2014.

[15] N. Gierasimczuk and D. de Jongh. On the complexity of conclusive update. *The Computer Journal*, 56(3):365–377, 2013.

[16] E. M. Gold. Limiting recursion. *JSL*, 30(1):28–48, 1965.

[17] E. M. Gold. Language identification in the limit. *Inform. Control*, 10:447–474, 1967.

[18] A. Grove. Two modellings for theory change. *JPL*, 17:157–170, 1988.

[19] K. T. Kelly. *The Logic of Reliable Inquiry.* Oxford University Press, Oxford, 1996.

[20] K. T. Kelly. Iterated belief revision, reliability, and inductive amnesia. *Erkenntnis*, 50:11–58, 1998.

[21] K. T. Kelly. The learning power of belief revision. In *Proc. of TARK'98*, pages 111–124. Morgan Kaufmann Publishers Inc., 1998.

[22] K. T. Kelly. Ockham's razor, truth, and information. In P. Adriaans and J. van Benthem, editors, *Handbook of the Philosophy of Information*, pages 321–359. Elsevier, 2008.

[23] K. T. Kelly. An erotetic theory of empirical simplicity and its connection with truth. Unpublished manuscript, 2011.

[24] K. T. Kelly. Notes on a general topological paradigm. *Unpublished manuscript*, 2015.

[25] K. T. Kelly and H. Lin. A simple theory of theoretical simplicity. Unpublished manuscript, 2011.

[26] K. T. Kelly, O. Schulte, and V. Hendricks. Reliable belief revision. In *Proc. of the 10th ICLMPS*, pages 383–398. Kluwer Academic Pub., 1995.

[27] D. Lewis. *Convention.* Cambridge: Harvard University Press, 1969.

[28] E. Martin and D. Osherson. Scientific discovery based on belief revision. *JSL*, 62(4):1352–1370, 1997.

[29] E. Martin and D. Osherson. *Elements of Scientific Inquiry.* MIT Press, Cambridge, MA, USA, 1998.

[30] H. Putnam. Trial and error predicates and the solution to a problem of Mostowski. *JSL*, 30(1):49–57, 1965.

[31] O. Schulte and C. Juhl. Topology as epistemology. *Monist*, 79(1):141–147, 1996.

[32] R. Solomonoff. A formal theory of inductive inference. *Inform. Control*, Part I, 7(1):1–22, 1964. Part II, 7(2):224–254, 1964, 1964.

[33] S. Vickers. *Topology Via Logic.* Cambridge University Press, 1996.

Bayesian Games with Intentions

Adam Bjorndahl
Carnegie Mellon University
Philosophy
Pittsburgh, PA 15213
abjorn@andrew.cmu.edu

Joseph Y. Halpern
Cornell University
Computer Science
Ithaca, NY 14853
halpern@cs.cornell.edu

Rafael Pass
Cornell University
Computer Science
Ithaca, NY 14853
rafael@cs.cornell.edu

ABSTRACT

We show that standard Bayesian games cannot represent the full spectrum of belief-dependent preferences. However, by introducing a fundamental distinction between *intended* and *actual* strategies, we remove this limitation. We define *Bayesian games with intentions*, generalizing both Bayesian games and psychological games [5], and prove that Nash equilibria in psychological games correspond to a special class of equilibria as defined in our setting.

1. INTRODUCTION

Type spaces were introduced by John Harsanyi [6] as a formal mechanism for modeling games of incomplete information where there is uncertainty about players' payoff functions. Broadly speaking, types are taken to encode payoff-relevant information, a typical example being how each participant values the items in an auction. An important feature of this formalism is that types also encode beliefs about types. Thus, a type encodes not only a player's beliefs about other players' payoff functions, but a whole *belief hierarchy*: a player's beliefs about other players' beliefs, their beliefs about other players' beliefs about other players' beliefs, and so on.

This latter point has been enthusiastically embraced by the epistemic game theory community, where type spaces have been co-opted for the analysis of games of *complete* information. In this context, types encode beliefs about the *strategies* used by other players in the game as well as their types. So again, types encode a belief hierarchy, but one that describes players' beliefs about other players' beliefs... about the other players' types and strategies. In this framework, one can determine whether a player is rational given her type and strategy; that is, whether her strategy is such that she is making a best response to her beliefs, as encoded by her type. Thus, rationality, common belief of rationality, and so on can be defined as events in the space of (profiles of) strategy-type pairs. This opens the way to epistemic analyses of solution concepts, among other applications [3]. In this setting, types do not encode payoff-relevant information; rather, they are simply a tool for describing belief hierarchies about other players' (types and) strategies.

By contrast, in a *Bayesian game*, types are payoff-relevant objects in that utility depends on them, though the payoff-relevant information they are taken to encode often includes such things as characteristics of the players (strength, work ethic, etc.), or more generally any relevant facts that may not be common knowledge. There is typically assumed to be a prior probability on types (indeed, often a common prior), so a type can still be viewed as encoding beliefs on other types in this setting (a type t encodes the probability obtained by conditioning the prior on t), and thus a belief hierarchy. However, the only aspect of this belief hierarchy that is typically used in Bayesian games is the first-order belief about other players' types (but not beliefs about beliefs, and so on), which is needed to defined a player's expected utility. Nonetheless, it is possible to leverage the fact that types encode beliefs to define Bayesian games in which players' preferences depend to some extent on the *beliefs* of their opponents (see Example 2.2). This observation is the point of departure for the present work.

The notion that a player's preferences might depend on the beliefs of her opponents (or on her own beliefs) is not new. *Psychological games* [1, 5] model phenomena like anger, surprise, and guilt by incorporating belief hierarchies directly into the domain of the utility functions. *Language-based games* [2] model similar belief-dependent preferences by defining utility over descriptions in a given language (in particular, a language that can express the players' beliefs). Types play no explicit role in either of these frameworks; on the other hand, the discussion above suggests that they may be naturally employed to accomplish many of the same modeling goals. Since Bayesian games and, more generally, type spaces have become cornerstones of modern game theory, if the modeling and analysis of psychological games could be carried out in this familiar framework, it would unify these paradigms and thereby amplify both the insights and the accessibility of the latter. In this paper, we provide an extension of Bayesian games that allows us to do just this.

There is an obvious obstruction to capturing general belief-dependent preferences using types in the standard way: types in Bayesian games encode beliefs about types, not about strategies. This severely limits the extent to which preferences over types can capture feelings like surprise or guilt, which are typically expressed by reference to beliefs about strategies (e.g., my opponent is surprised if I do not play the strategy that she was expecting me to play). It may seem that there is a simple solution to this problem: allow types to encode beliefs about strategies. But doing this leads to difficulties in the definition of *Bayesian Nash equilibrium*, the standard solution concept in Bayesian games; this notion depends on being able to freely associate strategies with types. In Section 2, we give the relevant definitions and make this issue precise.

In Section 3, we develop a modification of the standard Bayesian setup where each player is associated with *two* strategies: an *intended* strategy that is determined by her type (and thus can be the object of beliefs), and an *actual* strategy that is independent of her type (as in standard Bayesian games). This gives us what we call *Bayesian games with intentions*. We define a solution concept for such games where we require that, in equilibrium, the actual and intended strategies are equal. As we show, under this requirement, equilibria do not always exist.

In Section 4, we show that psychological games can be embedded in our framework. Moreover, we show that the notion of Nash equilibrium for psychological games defined by Geanakoplos, Pearce, and Stachetti [5] (hereafter GPS) corresponds to a special case of our own notion of equilibrium. Thus, we realize all the advantages of psychological games in an arguably simpler, better understood setting. We do not require complicated beliefs hierarchies; these are implicitly encoded by types.

The advantages of distinguishing actual from intended strategies go well beyond psychological games. As we show in the full paper, intended strategies can be fruitfully interpreted as *reference points* in the style of prospect theory [7]. One of the central insights of prospect theory is that the subjective value of an outcome can depend, at least in part, on how that outcome compares to some "reference level"; for example, whether it is viewed as a relative gain or loss. The intended/actual distinction naturally implements the needed comparison between "real" and "reference" outcomes. Using this insight, we show that *reference-dependent preferences*, as defined by Kőszegi and Rabin [8], can be captured using Bayesian games with intentions.

2. BAYESIAN GAMES

2.1 Definition

A *Bayesian game* is a model of strategic interaction among players whose preferences can depend on factors beyond the strategies they choose to play. These factors are often taken to be characteristics of the players themselves, such as whether they are industrious or lazy, how strong they are, or how they value certain objects. Such characteristics can be relevant in a variety of contexts: a job interview, a fight, an auction, etc.

A *type* of player i is often construed as encoding precisely such characteristics. More generally, however, types can be viewed as encoding any kind of information about the world that might be payoff-relevant. For example, the resolution of a battle between two armies may depend not only on what maneuvers they each perform, but also on how large or well-trained they were to begin with, or the kind of terrain they engage on. Decision-making in such an environment therefore requires a representation of the players' uncertainty regarding these variables.

We now give a definition of Bayesian games that is somewhat more general than the standard definition. This will make it easier for us to develop the extension to Bayesian games with intentions. We explain the differences after we give the definition.

Fix a set of *players*, $N = \{1, \dots, n\}$. A **Bayesian game (over** N**)** is a tuple $\mathcal{B} = (\Omega, (\Sigma_i, T_i, \tau_i, p_i, u_i)_{i \in N})$ where

- Ω is the measurable space of *states of nature*;
- Σ_i is the set of *strategies available to player* i;
- T_i is the set of *types of player* i;
- $\tau_i : \Omega \to T_i$ is *player* i*'s signal function*;
- $p_i : T_i \to \Delta(\Omega)$ associates with each type t_i of player i a probability measure $p_i(t_i)$ on Ω with $p_i(t_i)(\tau_i^{-1}(t_i)) = 1$, representing *type* t_i *of player* i*'s beliefs* about the state of nature;[1]
- $u_i : \Sigma \times \Omega \to \mathbb{R}$ is *player* i*'s utility function.*[2]

As we said above, this definition of a Bayesian game is more general than what is presented in much (though not all) of the literature. There are two main differences. First, we take utility to be defined over strategies and *states of nature*, rather than over strategies and types (cf. [9] for a similar definition). This captures the intuition that what is really payoff-relevant is *the way the world is*, and types simply capture the players' imperfect knowledge of this. Since the type signal function profile $(\tau_1, \dots, \tau_n)$ associates with each world a type profile, utilities can depend on players' types. Of course, we can always restrict attention to the special case where $\Omega = T$ and where $\tau_i : T \to T_i$ is the ith projection function; this is called the *reduced form*, and it accords with a common conception of types as encoding all payoff-relevant information aside from strategy choices (cf. [4]).

The second respect in which this definition is more general than is standard is in the association of an *arbitrary* probability measure $p_i(t_i)$ to each type t_i. It is typically assumed instead that for each player i there is some fixed probability measure $\pi_i \in \Delta(\Omega)$ representing her "prior beliefs" about the state of nature, and $p_i(t_i)$ is obtained by conditioning these prior beliefs on the "private information" t_i (or, more precisely, on the event $\tau_i^{-1}(t_i)$).[3] When $\pi_1 = \pi_2 = \cdots = \pi_n$, we say that the players have a *common prior*; this condition is also frequently assumed in the literature. We adopt the more general setup because it accords with a standard presentation of type spaces as employed for the epistemic analysis of games of complete information [3], thus making it easier for us to relate our approach to epistemic game theory.

The requirement that $p_i(t_i)(\tau_i^{-1}(t_i)) = 1$ amounts to assuming that each player is sure of her own type (and hence, her beliefs); that is, in each state $\omega \in \Omega$, each player i knows that the true state is among those where she is of type $t_i = \tau_i(\omega)$, which is exactly the set $\tau_i^{-1}(t_i)$.

2.2 Examples

It will be helpful to briefly consider two simple examples of Bayesian games, one standard and one a bit less so.

Example 2.1. First consider a simplified auction scenario where each participant $i \in N$ must submit a bid $\sigma_i \in \Sigma_i =$

[1] As usual, we denote by $\Delta(X)$ the set of probability measures on the measurable space X. To streamline the presentation, we suppress measurability assumptions here and elsewhere in the paper.

[2] Given a collection $(X_i)_{i \in N}$ indexed by N, we adopt the usual convention of denoting by X the product $\prod_{i \in N} X_i$ and by X_{-i} the product $\prod_{j \neq i} X_j$.

[3] To ensure this is well-defined, it is also typically assumed that none of player i's types are null with respect to π_i; that is, for all $t_i \in T_i$, $\pi_i(\tau_i^{-1}(t_i)) > 0$.

$\mathbb{R}^+$ for a given item. Types here are conceptualized as encoding valuations of the item up for auction: for each $t_i \in T_i$, let $v(t_i) \in \mathbb{R}^+$ represent how much player i thinks the item is worth, and define player i's utility $u_i : \Sigma \times T$ by

$$u_i(\sigma, t) = \begin{cases} v(t_i) - \sigma_i & \text{if } \sigma_i = \max_{j \in N} \sigma_j \\ 0 & \text{otherwise.} \end{cases}$$

Thus, a player's payoff is 0 if she does not submit the highest bid, and otherwise is equal to her valuation of the item less her bid (for simplicity, this model assumes that in the event of a tie, every top-bidding player gets the item). Note that the state space here is implicitly taken to be identical to the space T of type profiles, that is, the game is presented in reduced form. A type t_i therefore tells us not only how valuable player i thinks the item is ($v(t_i)$), but also what beliefs $p_i(t_i) \in \Delta(T)$ player i has about how the *other* players value the item (and what beliefs they have about *their* opponents, and so on). The condition that $p_i(t_i)(\tau^{-1}(t_i)) = 1$ then simply amounts to the assumption that each player is sure of her own valuation (as well as her beliefs about other players' types). □

Example 2.2. Next we consider an example where the Bayesian framework is leveraged to model a player whose preferences depend on the beliefs of her opponent. Consider a game where the players are students in a class, with player 1 having just been called upon by the instructor to answer a yes/no question. Assume for simplicity that $N = \{1, 2\}$, $\Sigma_1 = \{\mathsf{yes}, \mathsf{no}, \mathsf{pass}\}$, and $\Sigma_2 = \{*\}$ (where $*$ denotes a vacuous move, so only player 1 has a real decision to make). Let $\Omega = \{w_y, w_n, v_y, v_n\}$, where, intuitively, states with the subscript y are states where "yes" is the correct answer, while states with the subscript n are states where "no" is the correct answer. Let $T_1 = \{t_1, t'_1\}$, $T_2 = \{t_2, t'_2, t''_2\}$, and define the signal functions by

$$\tau_1(w_y) = \tau_1(w_n) = t_1, \; \tau_1(v_y) = \tau_1(v_n) = t'_1, \text{ and}$$
$$\tau_2(w_y) = \tau_2(w_n) = t_2 \text{ and } \tau_2(v_y) = t'_2 \text{ and } \tau_2(v_n) = t''_2.$$

Finally, assume that all of the subjective probability measures arise by conditioning a common prior $\pi \in \Delta(\Omega)$ on the type of the player in question; assume further that π is the uniform distribution. It follows that in each state, player 1 is unsure of the correct answer. On the other hand, while in states w_y and w_n, player 2 is also unsure of the correct answer, in states v_y and v_n, player 2 knows the correct answer. Moreover, in states w_y and w_n, player 1 is sure that player 2 does not know the correct answer, whereas in states v_y and v_n, player 1 is sure that player 2 *does* know the correct answer (despite not knowing it himself). We can therefore use this framework to encode the following (quite plausible) preferences for player 1: guessing the answer is preferable to passing provided player 2 does not know the right answer, but passing is better than guessing otherwise. Set

$$u_1(\mathsf{yes}, w_y) = u_1(\mathsf{yes}, v_y) = u_1(\mathsf{no}, w_n) = u_1(\mathsf{no}, v_n) = 5,$$

representing a good payoff for answering correctly; set

$$u_1(\mathsf{pass}, x) = -2 \text{ for all } x \in \Omega,$$

representing a small penalty for passing regardless of what the correct answer is; finally, set

$$u_1(\mathsf{yes}, w_n) = u_1(\mathsf{no}, w_y) = -5 \text{ and}$$
$$u_1(\mathsf{yes}, v_n) = u_1(\mathsf{no}, v_y) = -15,$$

representing a penalty for getting the wrong answer that is substantially worse in states where player 2 knows the correct answer.

It is easy to check that if player 1 considers w_y and w_n to be equally probable, then her expected utility for randomly guessing the answer is 0, which is strictly better than passing (passing, of course, always yields an expected utility of -2). By contrast, if player 1 considers v_y and v_n to be equally probable, then her expected utility for randomly guessing is -5, which is strictly worse than passing. In short, player 1's decision depends on what she believes about the beliefs of player 2. □

Example 2.2 captures what might be thought of as *embarrassment aversion*, which is a species of belief-dependent preference: player 1's preferences depend on what player 2 believes. It is worth being explicit about the conditions that make this possible:

C1. States in Ω encode a certain piece of information I (in this case, whether the correct answer to the given question is "yes" or "no").

C2. Types encode beliefs about states.

C3. Utility depends on types.

From C1–C3, we can conclude that preferences can depend on what the players believe about I.

Not all kinds of belief-dependent preferences can be captured in the Bayesian framework. Suppose, for example, that the goal of player 1 is to surprise her opponent by playing an unexpected strategy. More precisely, suppose that $\Sigma_1 = \{\sigma_1, \sigma'_1\}$ and we wish to define u_1 in such a way that player 1 prefers to play σ_1 if and only if player 2 believes he will play σ'_1. In contrast to Example 2.2, this scenario cannot be represented with a Bayesian game for the following simple reason: *states do not encode strategies.* In other words, condition C1 is not satisfied if we take I to be player 1's strategy. Therefore, types cannot encode such beliefs about strategies, so utility cannot be defined in a way that depends on such beliefs.

This suggests an obvious generalization of the Bayesian setting, namely, encoding strategies in states. Indeed, this is the idea we explore in this paper; however, it is not quite as straightforward a maneuver as it might appear, primarily due to its interaction with the mechanics of *Bayesian Nash equilibrium.*

2.3 Equilibrium

Part of the value of Bayesian games lies in the fact that a generalized notion of Nash equilibrium can be defined in this framework, for which the following notion plays a crucial role: a *behaviour rule* for player i is a function $\beta_i : T_i \to \Sigma_i$. In Bayesian games, we talk about behaviour rule profiles being in equilibrium, just as in normal-form games, we talk about strategy profiles being in equilibrium. Intuitively, $\beta_i(t_i)$ represents the strategy that type t_i of player i is playing, so a player's strategy depends on her type.

From a technical standpoint, behaviour rules are important because they allow us to associate a payoff for each player with each *state*, rather than strategy-state pairs. Since types encode beliefs about states, this yields a notion of expected utility for each type. A Bayesian Nash equilibrium is then defined to be a profile of behaviour rules such that each type is maximizing its own expected utility.

More precisely, observe that via the signal functions τ_i, a behaviour rule β_i associates with each state ω the strategy $\beta_i(\tau_i(\omega))$. Thus, a profile β of behaviour rules defines an *induced utility function* $u_i^\beta : \Omega \to \mathbb{R}$ as follows:

$$u_i^\beta(\omega) = u_i((\beta_j(\tau_j(\omega)))_{j\in N}, \omega).$$

The beliefs $p_i(t_i)$ then define the *expected utility* for each type: let $E_{t_i}(\beta)$ denote the expected value of u_i^β with respect to $p_i(t_i)$. Denote by B_i the set of all behaviour rules for player i. A behaviour rule β_i is a **best response to** β_{-i} if, for each $t_i \in T_i$, β_i maximizes E_{t_i}:

$$(\forall \beta_i' \in B_i)(E_{t_i}(\beta_i, \beta_{-i}) \geq E_{t_i}(\beta_i', \beta_{-i})).$$

Finally, a **Bayesian Nash equilibrium** of the Bayesian game $\mathcal{B}$ is a profile of behaviour rules β such that, for each $i \in N$, β_i is a best response to β_{-i}. A (mixed) Bayesian Nash equilibrium is guaranteed to exist when the strategy and types spaces are finite (see [10] for a more general characterization of when an equilibrium exists).

3. INTENTION

3.1 Definition

Behaviour rules map types to strategies, but the underlying model does not enforce any relationship between types and strategies (or between states and strategies). Thus, behaviour rules do not provide a mechanism satisfying condition C1 with I taken to be a player's strategies, so they do not allow us to express preferences that depend on beliefs about strategies. In order to express such preferences, we must associate strategies with states in the model itself. Note that once we do this, utility functions depend on strategies in two ways. Specifically, since u_i is defined on the cross product $\Sigma\times\Omega$, players' preferences depend on strategies both directly (corresponding to the strategy-profile component of u_i's input) and as encoded in states (the second component of u_i's input). To keep track of this distinction, we call these *actual* and *intended* strategies, respectively.

Formally, a **Bayesian game with instantiated intentions** (BGII) is a tuple $\mathcal{I} = (\Omega, (\Sigma_i, T_i, \tau_i, s_i, p_i, u_i)_{i\in N})$, where $s_i : T_i \to \Sigma_i$ is *player i's intention function* and the remaining components are defined as in a Bayesian game. (The reason for this terminological mouthful will become clear in Section 3.3, where we define *Bayesian games with intentions.*) Each s_i associates with each type t_i of player i an *intended strategy* $s_i(t_i)$. Intuitively, we might think of $s_i(t_i)$ as the strategy that a player of type t_i "intends" or "is planning" to play (though may ultimately decide not to); alternatively, it might be conceptualized as the "default" strategy for that type; it might even be viewed as the "stereotypical" strategy employed by players of type t_i. The former interpretation may be appropriate in a situation where we want to think of self-control; for example, a player who intends to exercise, but actually does not. The latter interpretation may be appropriate if we think about voting. Wealthy people in Connecticut typically vote Republican, but a particular player i who is wealthy and lives in Connecticut (this information is encoded in her type) votes Democrat.

We associate intended strategies with types rather than directly with states by analogy to behaviour rules, in keeping with the modeling paradigm where the personal characteristics of a player—including her beliefs, decisions, *and intentions*—are entirely captured by her type. Nonetheless, the composition $s_i \circ \tau_i : \Omega \to \Sigma_i$ does associate strategies with states and so satisfies condition C1 (again, with I being a player's strategy); thus, players can have beliefs about strategies. This, in turn, allows us to define utility so as to capture preferences that depend on beliefs about strategies.

3.2 Examples

The presentation of a BGII is made clearer by introducing the following notation for the set of states where player i intends to play σ_i:

$$[\![\sigma_i]\!] = (s_i \circ \tau_i)^{-1}(\sigma_i) = \{\omega \in \Omega \,:\, s_i(\tau_i(\omega)) = \sigma_i\}.$$

Example 3.1. Consider a 2-player game in which player 1's goal is to surprise her opponent. We take player 2 to be surprised if his beliefs about what player 1 intends to play are dramatically different from what player 1 actually plays. For definiteness, we take "dramatically different" to mean that his beliefs about player 1's intended strategy ascribe probability 0 to player 1's actual strategy. Thus, we define player 1's utility function as follows:

$$u_1(\sigma, \omega) = \begin{cases} 1 & \text{if } p_2(\tau_2(\omega))([\![\sigma_1]\!]) = 0 \\ 0 & \text{otherwise.} \end{cases}$$

(Recall that $p_2(\tau_2(\omega))$ is a measure on states, which is why we apply it to $\tau_1^{-1}(s_1^{-1}(\sigma_1))$, that is, the set of states ω where player 1's intended strategy, $s_1(\tau_1(\omega))$, is equal to σ_1.) □

Example 3.2. Next we consider an example introduced by GPS [5] called *the bravery game*. This is a 2-player scenario in which player 1 has the only real decision to make: he must choose whether to take a *bold* action or a *timid* action, so $\Sigma_1 = \{\mathsf{bold}, \mathsf{timid}\}$ (and $\Sigma_2 = \{*\}$). The crux of the game is the psychological factor, described by GPS as follows: player 1 prefers "to be timid rather than bold, unless he thinks his friends expect him to be bold, in which case he prefers not to disappoint them" [5]. It is also stipulated that player 2 prefers player 1 to be bold, and also prefers to think of him as bold. Define $q : T \to [0, 1]$ to be the degree of belief that type t_2 of player 2 has in player 1 being bold:

$$q(t) = p(t_2)([\![\mathsf{bold}]\!]).$$

Define $\tilde{q} : T \to [0, 1]$ to be type t_1 of player 1's expectation of this degree of belief:

$$\tilde{q}(t) = E_{t_1}(q),$$

where $E_{t_i}(f)$ denotes the expected value of f with respect to the measure $p(t_i)$. We can then represent the players' preferences in a reduced-form BGII as follows:

$$u_1(\sigma, t) = \begin{cases} 2 - \tilde{q}(t) & \text{if } \sigma_1(t_1) = \mathsf{bold} \\ 3(1 - \tilde{q}(t)) & \text{if } \sigma_1(t_1) = \mathsf{timid}, \end{cases}$$

$$u_2(\sigma, t) = \begin{cases} 2(1 + q(t)) & \text{if } \sigma_1(t_1) = \mathsf{bold} \\ 1 - q(t) & \text{if } \sigma_1(t_1) = \mathsf{timid}. \end{cases}$$

This representation closely parallels that given in [5], in which q and $\tilde{q}$ are understood not as functions of types, but (implicitly) as functions of belief hierarchies.[4] But this makes no difference to the preferences this game encodes. For example, it is easy to see that player 2 prefers player 1 to be bold, and all the more so when q is high—that is, all the more so when she believes with high probability that he will be bold.[5] Similarly, one can check that player 1 prefers to be timid provided that $\tilde{q}(t) < \frac{1}{2}$; in other words, provided that his expectation of his opponent's degree of belief in him being bold is sufficiently low.

Why not define player 1's preferences directly in terms of the beliefs of his opponent, rather than his expectation of these beliefs? GPS cannot do so because of a technical limitation of the framework as developed in [5]; specifically, that a player's utility can depend only on *her own* beliefs. Battigalli and Dufwenberg [1] correct this deficiency. BGIIs do not encounter such limitations in the first place. In particular, it is easy enough to redefine player 1's utility as follows:

$$u_1'(\sigma, t) = \begin{cases} 2 - q(t) & \text{if } \sigma_1(t_1) = \mathsf{bold} \\ 3(1 - q(t)) & \text{if } \sigma_1(t_1) = \mathsf{timid}. \end{cases}$$

In this case, we find that player 1 prefers to be timid provided $q(t) < \frac{1}{2}$, or in other words, provided that his opponent's degree of belief in him being bold is sufficiently low. □

Observe that in neither of the preceding examples did we provide a concrete BGII, in that we did not explicitly define the type spaces, the intention functions, and so on. Instead, we offered general recipes for implementing certain belief-dependent preferences (e.g., to surprise, to live up to expectations, etc.) in arbitrary BGIIs. Particular choices of type spaces and intention functions do play an important role in equilibrium analyses; however, as illustrated by the preceding two examples, at the modeling stage they need not be provided up front.

3.3 Equilibrium

We now define a notion of equilibrium for this setting. As a first step towards this definition, given a BGII $\mathcal{I}$, we say that a profile of behaviour rules β is an **equilibrium of $\mathcal{I}$** provided:

(1) β is a Bayesian Nash equilibrium of the underlying Bayesian game: that is, each β_i is a best response to β_{-i} in precisely the sense defined in Section 2.3;

(2) for each player $i \in N$, $\beta_i = s_i$.

This definition, and in particular condition (2), embodies the conception of equilibrium as a steady state of play where each player has correct beliefs about her opponents (and is best responding to those beliefs). In a BGII, beliefs about the strategies of one's opponents are beliefs about intended strategies (although, in equilibrium, a player will also have beliefs about actual strategies). On the other hand, since behavior rules associate strategies with types and players have beliefs over types, behaviour rules also induce beliefs about strategies; in our terminology, these are beliefs about actual strategies. Condition (2) implies that these two beliefs coincide in equilibrium; in equilibrium, each type of each player actually plays the strategy she intended to play (which is exactly the strategy her opponents expected her to play).

Does condition (2) collapse the distinction between intended and actual strategies, thereby returning us to the classical setting? It does not. First, in a standard Bayesian game we could not even write down a model where players' preferences depended on beliefs about strategies. In addition, although we demand that intended and actual strategies coincide in equilibrium, this restriction *does not apply to the evaluation of best responses*. Recall that β_i is a best response to β_{-i} if and only if

$$(\forall \beta_i' \in B_i)(E_{t_i}(\beta_i, \beta_{-i}) \geq E_{t_i}(\beta_i', \beta_{-i})).$$

Crucially, β_i' need not be equal to s_i. In other words, for β_i to count as a best response, it must be at least as good as all other behaviour rules, including those that recommend playing a strategy distinct from that specified by s_i.

Example 3.3. Consider a 2-player reduced-form BGII with $\Sigma_1 = \{a, b\}$, $\Sigma_2 = \{*\}$, $T_1 = \{x, x'\}$, and $T_2 = \{y, y'\}$, and where

$$p_1(x)(\{y\}) = p_1(x')(\{y'\}) = p_2(y)(\{x'\}) = p_2(y')(\{x\}) = 1.$$

Let u_1 be defined as in Example 3.1, encoding player 1's desire to surprise her opponent:

$$u_1(\sigma_1, *, t) = \begin{cases} 1 & \text{if } p_2(t_2)(\llbracket \sigma_1 \rrbracket) = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that $s_1(x) = s_1(x') = a$. Then, of course, $p_2(y)(\llbracket a \rrbracket) = p_2(y')(\llbracket a \rrbracket) = 1$, and likewise $p_2(y)(\llbracket b \rrbracket) = p_2(y')(\llbracket b \rrbracket) = 0$. It follows immediately that the expected utility of playing a for either type of player 1 is equal to 0 (since player 1 is sure that this will not surprise her opponent), whereas the expected utility of playing b for either type of player 1 is equal to 1 (since, in this case, player 1 is sure that this *will* surprise her opponent). In particular, if $\beta_1 = s_1$, then β_1 is not a best response. Thus, this particular BGII admits no equilibrium.

Now suppose that $s_1(x) = a$ and $s_1(x') = b$. This is, of course, a different BGII from the one considered in the previous paragraph, but it differs only in the specification of player 1's intentions. Moreover, in this BGII it is not hard to check that $\beta_1 = s_1$ is a best response and therefore constitutes an equilibrium: type x is sure that player 2 is of type y; therefore, type x is sure that player 2 is sure that player 1 is of type x', and so is playing b; thus, a is a best response for x, since x is sure that it will surprise her opponent; a similar argument shows that b is a best response for x'. □

Example 3.3 demonstrates that the notion of best response in a BGII—and therefore the notion of equilibrium—can be sensitive to states of play where players are *not* playing their

[4] Additionally, GPS give the value of q, not by the probability that player 2 assigns to player 1 being bold, but by player 2's *expectation* of the probability p with which player 1 decides to be bold. We forgo this subtlety for the time being.

[5] It is not quite clear why GPS define player 2's payoff in the event that player 1 is timid to be $1 - q(t)$ rather than $1+q(t)$. This latter value preserves the preferences described while avoiding the implication that, assuming that player 1 will be timid, player 2 also prefers to *believe* that he will be timid—this stands in opposition to the stipulation that player 2 prefers to think of her opponent as bold.

intended strategies. But it also illustrates the pivotal role of the intention functions s_i in determining the existence of an equilibrium. Indeed, condition (2) implies that if a given BGII $\mathcal{I}$ has an equilibrium at all, it is unique and equal to s. This suggests that BGIIs are not at the right "resolution" for equilibrium analysis, since they come already equipped with a unique candidate for equilibrium. Thus, rather than restricting attention to a single BGII, where the intention function is specified and hard-coded into the model, we consider a more general model, where the intention function is not specified, but still affects the utility. This is parallel to the role of strategies in standard games, which are not hard-coded into the model, but of course the utility function still depends on them. Essentially, we are moving the intention function from the model to the utility function. As we shall see, our earlier examples of BGIIs can be easily interpreted as models in this more general sense.

In order to make this precise, we must first formally define utility functions that take as arguments intention functions. More precisely, taking $\Sigma^T = \Sigma_1^{T_1} \times \cdots \times \Sigma_n^{T_n}$ (so that Σ^T is the set of intention function profiles), an *explicit utility function* is a map $\tilde{u}_i : \Sigma \times \Omega \times \Sigma^T \to \mathbb{R}$; these are just like the utility functions in a BGII except they explicitly take as input the associations between types and strategies provided by intention functions. A **Bayesian game with intentions** (BGI) is a tuple $\tilde{\mathcal{I}} = (\Omega, (\Sigma_i, T_i, \tau_i, p_i, \tilde{u}_i)_{i \in N})$, where the components are defined just as they are in a Bayesian game, except that the functions $\tilde{u}_i$ are explicit utility functions. We emphasize that a BGI, unlike a BGII, does not include players' intention functions among its components; instead, these functions show up as arguments in the (explicit) utility functions.

It is easy to see that all the examples of BGIIs that we have considered so far can be naturally converted to BGIs. For example, the utility function $u_1(\sigma, t)$ in Example 3.1 becomes $\tilde{u}_i(\sigma, t, s)$. The definition of $\tilde{u}_i(\sigma, t, s)$ looks identical to that of $u_1(\sigma, t)$; the additional argument s is needed to define $[\![\sigma_1]\!]$.

A BGI induces a natural map from intention functions to BGIIs: given $\tilde{\mathcal{I}} = (\Omega, (\Sigma_i, T_i, \tau_i, p_i, \tilde{u}_i)_{i \in N})$ and functions $s_i : T_i \to \Sigma_i$, let

$$\tilde{\mathcal{I}}(s_1, \ldots, s_n) = (\Omega, (\Sigma_i, T_i, \tau_i, s_i, p_i, u_i)_{i \in N}),$$

where $u_i : \Sigma \times \Omega \to \mathbb{R}$ is defined by

$$u_i(\sigma, \omega) = \tilde{u}_i(\sigma, \omega, s_1, \ldots, s_n).$$

Clearly $\tilde{\mathcal{I}}(s_1, \ldots, s_n)$ is a BGII; we call it an **instantiation of $\tilde{\mathcal{I}}$**. We then define an **equilibrium of $\tilde{\mathcal{I}}$** to be a profile of behaviour rules β that is an equilibrium of the corresponding instantiation $\tilde{\mathcal{I}}(\beta)$. Here we make implicit use of the fact that both behaviour rules and intention functions are functions from types to strategies. Indeed, the profile β plays two roles: first, it is used to determine the intentions of the players; then, in the context of the instantiated BGI with these fixed intentions, we evaluate whether each $\beta_i(t_i)$ is a best response, just as in the definition of equilibrium for a standard Bayesian game.

Is this a reasonable notion of equilibrium? As we observed above, in a BGII, the only possible equilibrium is "built in" to the model in the form of the intention functions. In particular, the only possible equilibrium for the instantiation $\tilde{\mathcal{I}}(\beta)$ is β itself. Of course, β is not necessarily an equilibrium of this game; however, by quantifying over β and considering the corresponding class of BGIIs (i.e., those obtained as instantiations of $\tilde{\mathcal{I}}$), we are essentially asking the question: "Is there a profile of intentions such that, assuming those intentions are common knowledge, no player prefers to deviate from their intention?" If so, that profile constitutes an equilibrium. This is a natural solution concept; in fact, as we show in Section 4, the notion of equilibrium proposed by GPS for psychological games is a special case of our definition.

Example 3.4. In light of these definitions, Example 3.3 can be viewed as first defining a BGI $\tilde{\mathcal{I}}$, and then considering two particular instantiations of it. The equilibrium observations made then amount to the following: the behaviour rule $\beta_1 \equiv a$ (i.e., the constant function a) is not an equilibrium of $\tilde{\mathcal{I}}$, but the behaviour rule β_1' that sends x to a and x' to b is. (As there is only ever one option for player 2's behaviour rule, namely $\beta_2 \equiv *$, we can safely neglect it.) □

Example 3.5. Consider again the bravery game of Example 3.2. Under any particular specification of state space and type spaces, this becomes a BGI $\tilde{\mathcal{I}}$. It is not difficult to see that each of the behaviour rules $\beta_1 \equiv \mathsf{timid}$ and $\beta_1' \equiv \mathsf{bold}$ is an equilibrium of $\tilde{\mathcal{I}}$. □

3.4 Existence

Are equilibria of BGIs guaranteed to exist? Not necessarily. At least one obstacle to existence lies in the specification of the underlying type space and the corresponding probability measures: as the following example shows, certain kinds of belief that are necessary for best-responses may be implicitly ruled out.

Example 3.6. Consider a 2-player reduced-form BGI $\tilde{\mathcal{I}}$ where $\Sigma_1 = \{a, b\}$, $\Sigma_2 = \{*\}$, $T_1 = \{x, x'\}$, and $T_2 = \{y, y'\}$, and where

$$p_1(x)(\{y\}) = p_1(x')(\{y'\}) = p_2(y)(\{x\}) = p_2(y')(\{x'\}) = 1.$$

Once again we consider a model where player 1 wishes to surprise her opponent, and so define u_1 as in Example 3.3:

$$u_1(\sigma_1, *, t) = \begin{cases} 1 & \text{if } p_2(t_2)([\![\sigma_1]\!]) = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that player 1 is certain that player 2 knows her type. It follows that no matter what her intentions are, player 2 knows them, and so (by definition of u_1), player 1 can always do better by deviating. In other words, no behaviour rule β_1 is an equilibrium of $\tilde{\mathcal{I}}(\beta_1)$ (since it is not a best response). It follows immediately that $\tilde{\mathcal{I}}$ admits no equilibria. □

This obstacle persists even if we extend our attention to mixed strategies. More precisely, consider the class of BGIIs where, for each player i, $\Sigma_i = \Delta(A_i)$ for some finite set A_i (the set of player i's *pure strategies*), and $u_i : \Sigma \times \Omega \to \mathbb{R}$ satisfies

$$u_i(\sigma_i, \sigma_{-i}, \omega) = \sum_{a_i \in A_i} \sigma_i(a_i) u_i(a_i, \sigma_{-i}, \omega).$$

In other words, player i's utility for playing σ_i is just the expected value of her utility for playing her various pure strategies with the probabilities given by σ_i. As is standard, we call elements of Σ_i *mixed strategies*, and the corresponding BGIIs *mixed-strategy BGIIs*. We can similarly define *mixed-strategy BGIs*. Note that in this context, since the intention functions s_i map into Σ_i, intended strategies are also mixed.

The next example shows that, in contrast to the classical setting, there are mixed-strategy BGIs with finite type spaces that admit no equilibria.

Example 3.7. Consider a 2-player reduced-form BGI where $\Sigma_1 = \Delta(\{a, b\})$, $\Sigma_2 = \{*\}$, $T_1 = \{x, x'\}$, and $T_2 = \{y, y'\}$, and where

$$p_1(x)(\{y\}) = p_1(x')(\{y'\}) = p_2(y)(\{x\}) = p_2(y')(\{x'\}) = 1.$$

Set

$$u_1(a, *, t) = \begin{cases} 1 & \text{if } p_2(t_2)(\llbracket a \rrbracket) < 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$u_1(b, *, t) = \begin{cases} 1 & \text{if } p_2(t_2)(\llbracket a \rrbracket) = 1 \\ 0 & \text{otherwise,} \end{cases}$$

and extend to all $\sigma_1 \in \Delta(\{a, b\})$ by taking expectation:

$$u_1(\sigma_1, *, t) = \sigma_1(a)u_1(a, *, t) + \sigma_1(b)u_1(b, *, t).$$

Note that, following standard conventions, here we identify the pure strategy a with the degenerate mixed strategy that places probability 1 on a; likewise for b. Thus, for example, the condition $p_2(t_2)(\llbracket a \rrbracket) < 1$ amounts to the following: "type t_2 is not absolutely certain that player 1 intends to play the pure strategy a", or equivalently, "type t_2 considers it possible that player 1 intends to play a mixed strategy that places positive weight on b". The preferences defined by u_1 can be roughly summarized as follows: "player 1 prefers to play a in the event that player 2 thinks she might place positive weight on b, and prefers to play b if player 2 is sure that she'll play a for sure".

This game admits no equilibria. To see this, suppose that β_1 were an equilibrium: that is, set player 1's intention function equal to β_1, and suppose that β_1 is an equilibrium of the resulting BGII.[6] First consider the case where $\beta_1(x) \in \Sigma_1$ satisfies $\beta_1(x)(b) > 0$. Then it follows that $p_2(y)(\llbracket a \rrbracket) = 0$ (i.e., type y is certain that player 1 is not playing the pure strategy a), and so, since type x is certain that player 2 is of type y, it follows by definition of u_1 that type x's best response is to play the pure strategy a. In particular, $\beta_1(x)$ is not a best response, so β_1 cannot constitute an equilibrium. Now consider the case where $\beta_1(x)(b) = 0$; in other words, $\beta_1(x)$ is the pure strategy a. Then we have $p_2(y)(\llbracket a \rrbracket) = 1$, from which it follows that type x's best response is to play the pure strategy b. Thus, once again, β_1 cannot constitute an equilibrium. □

4. PSYCHOLOGICAL GAMES

Psychological games can be captured in our framework. A psychological game $\mathcal{P}$ consists of a finite set of players N, together with mixed strategies Σ_i and utility functions $v_i : \bar{B}_i \times \Sigma \to \mathbb{R}$ for each player i, where $\bar{B}_i$ denotes the set of "collectively coherent" belief hierarchies for player i. Somewhat more precisely, an element $b_i \in \bar{B}_i$ is an infinite sequence of probability measures $(b_i^1, b_i^2, \ldots)$ where $b_i^1 \in \Delta(\Sigma_{-i})$ is player i's *first-order beliefs*, b_i^2 is player i's *second-order beliefs* (i.e., roughly speaking, her beliefs about the beliefs of her opponents), and so on, such that the beliefs in this sequence satisfy certain technical conditions (roughly speaking, lower-order beliefs must agree with the appropriate marginals of higher-order beliefs, and this agreement must be common knowledge); see the full paper for the complete definition.

Given a mixed-strategy BGII $\mathcal{I}$ and a type $t_i \in T_i$, we can define the first-order beliefs associated with t_i by

$$\varphi_i^1(t_i) = (s_{-i})_*(p_i(t_i));$$

that is, the pushforward of $p_i(t_i)$ from Ω to Σ_{-i} by s_{-i}. Note that, in our terminology, these are beliefs about *intended* strategies. The kth-order beliefs associated with t_i, denoted $\varphi_i^k(t_i)$, can be defined inductively in a similar fashion; it is then straightforward to show that the sequence

$$\varphi_i(t_i) = (\varphi_i^1(t_i), \varphi_i^2(t_i), \ldots)$$

is collectively coherent, and thus $\varphi_i : T_i \to \bar{B}_i$ (see the full paper).

This correspondence provides a natural notion of equivalence between psychological games and BGIIs with respect to the psychological preferences expressed in the former, namely,

$$\forall i \in N\ \forall \sigma \in \Sigma\, \forall \omega \in \Omega(u_i(\sigma, \omega) = v_i(\varphi_i(\tau_i(\omega)), \sigma)).$$

When a BGII $\mathcal{I}$ satisfies this condition with respect to a psychological game $\mathcal{P}$, we say that $\mathcal{I}$ and $\mathcal{P}$ are **preference-equivalent**.

The notion of preference-equivalence lifts naturally to BGIs. Observe that the functions φ_i^k depend on the profile of intention functions s; being explicit about this dependence, we write $\varphi_i^k(t_i; s)$ rather than $\varphi_i^k(t_i)$; we then say that $\tilde{\mathcal{I}}$ and $\mathcal{P}$ are preference-equivalent provided that

$$\forall i \in N\, \forall \sigma \in \Sigma\, \forall \omega \in \Omega\, \forall s \in \Sigma^T(\tilde{u}_i(\sigma, \omega, s) = v_i(\varphi_i(\tau_i(\omega); s), \sigma)).$$

It is easy to see that, given a psychological game $\mathcal{P}$, we can obtain a preference-equivalent BGI $\tilde{\mathcal{I}}$ simply by taking the above condition as the *definition* of the utility functions $\tilde{u}_i$. Thus, we have the following:

Proposition 4.1. *For every psychological game there exists a preference-equivalent BGI.*

Note that even very simple BGIs (i.e., those with very small type/state spaces) can be preference-equivalent to psychological games; indeed, it is sufficient for the utility functions $\tilde{u}_i$ to be of the form

$$\tilde{u}_i(\sigma, \omega, s) = f(\varphi_i(\tau_i(\omega); s), \sigma),$$

so that utility depends on states only to the extent that states encode belief hierarchies. In particular, although the utility functions in a psychological game have uncountable domains (since they apply to all possible belief hierarchies), a BGI $\tilde{\mathcal{I}}$ can be preference-equivalent to a psychological game $\mathcal{P}$ even if $\tilde{\mathcal{I}}$ has only finitely many states, since all that matters is that the utility functions of $\tilde{\mathcal{I}}$ agree with the utilitiy functions of $\mathcal{P}$ on the belief hierarchies encoded by the states of $\tilde{\mathcal{I}}$. Given a psychological game, we can construct a preference-equivalent BGI with type spaces rich enough that each φ_i is surjective: in other words, every belief hierarchy is realized by some type. However, in order

[6] As before, we ignore player 2's behaviour since he has no choices to make.

to capture *equilibrium* behaviour, such richness turns out to be superfluous. We now show how the notion of equilibrium defined by GPS for psychological games can be recovered as equilibria in our setting.

Given $\sigma \in \Sigma$, let $\chi_i(\sigma) \in \bar{B}_i$ denote the unique belief hierarchy for player i corresponding to common belief in σ. A *psychological Nash equilibrium* of $\mathcal{P}$ is a strategy profile σ such that, for each player i, σ_i maximizes the function

$$\sigma_i' \mapsto v_i(\chi_i(\sigma), \sigma_i', \sigma_{-i}).$$

In particular, to check whether σ constitutes a psychological Nash equilibrium, the only relevant belief hierarchies are those corresponding to common belief of σ. This, in essense, is the reason we do not need rich type spaces in BGIs to detect such equilibria.

THEOREM 4.2. *If $\mathcal{P}$ and $\tilde{\mathcal{I}}$ are preference-equivalent, then σ is a psychological Nash equilibrium of $\mathcal{P}$ if and only if the profile of (constant) behaviour rules β for which $\beta_i \equiv \sigma_i$ is an equilibrium of $\tilde{\mathcal{I}}$.*

PROOF. When β is the profile of behaviour rules described in this theorem, the corresponding instantiation $\tilde{\mathcal{I}}(\beta)$ has the property that, for each type t_i, $\varphi_i(t_i) = \chi_i(\sigma)$. The rest of the proof is essentially just unwinding definitions; see the full paper for details. □

Theorem 4.2 shows that equilibrium analysis in psychological games does not depend on the full space of belief hierarchies; it can be captured by particularly simple BGIs. It also establishes an equivalence between psychological Nash equilibria and a certain restricted class of equilibria in BGIs; namely, those consisting of constant behaviour rules. This restriction is not surprising in light of the fact that psychological games do not model strategies as functions of types, while BGIs do. Thus, BGIs are not merely recapitulations of the GPS framework: they are a common generalization of psychological games and Bayesian games.

5. CONCLUSION

We have introduced BGIs, Bayesian games with intentions, which generalize Bayesian games and psychological games in a natural way. We believe that BGIs will prove much easier to deal with than psychological games, while allowing greater flexibility.

When do equilibria exist? While Theorem 4.2 provides sufficient conditions for the existence of equilibria in BGIs, they are certainly not necessary conditions. We can show, for example, that there are BGIs that admit only equilibria in which no behaviour rule is constant. Formulating more general conditions sufficient for existence is the subject of ongoing work.

Perhaps the most exciting prospect for future research lies in leveraging the distinction between actual and intended strategies. As we show in the full paper, this distinction can be used to implement Köszegi and Rabin's [8] model of reference-dependent preferences; we believe that it will have other uses as well, and perhaps lead to new insights into solution concepts.

Acknowledgements

We are indebted to Aviad Heifetz for asking the question that first prompted us to explore the notion of extending Bayesian games to capture belief-dependent preferences. Halpern is supported in part by NSF grants IIS-0911036 and CCF-1214844, by ARO grant W911NF-14-1-0017, and by the Multidisciplinary University Research Initiative (MURI) program administered by the AFOSR under grant FA9550-12-1-0040. Pass is supported in part by a Microsoft Research Faculty Fellowship, NSF CAREER Award CCF-0746990, AFOSR Award FA9550-08-1-0197, and BSF Grant 2006317.

6. REFERENCES

[1] P. Battigalli and M. Dufwenberg. Dynamic psychological games. *Journal of Economic Theory*, 144:1–35, 2009.

[2] A. Bjorndahl, J. Y. Halpern, and R. Pass. Language-based games. In *Theoretical Aspects of Rationality and Knowledge: Proc. Fourteenth Conference (TARK 2013)*, pages 39–48, 2013.

[3] E. Dekel and M. Siniscalchi. Epistemic game theory. In H. P. Young and S. Zamir, editors, *Handbook of Game Theory with Economic Applications, Volume 4*, pages 619–702. North-Holland, 2015.

[4] D. Fudenberg and J. Tirole. *Game Theory.* MIT Press, Cambridge, Mass., 1991.

[5] J. Geanakoplos, D. Pearce, and E. Stacchetti. Psychological games and sequential rationality. *Games and Economic Behavior*, 1(1):60–80, 1989.

[6] J. Harsanyi. Games with incomplete information played by 'Bayesian' players, parts I–III. *Management Science*, 14:159–182, 320–334, 486–502, 1968.

[7] D. Kahneman and A. Tversky. Prospect theory, an analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

[8] B. Kőszegi and M. Rabin. A model of reference-dependent preferences. *The Quarterly Journal of Economics*, CXXI:1133–1165, 2006.

[9] M. J. Osborne and A. Rubinstein. *A Course in Game Theory.* MIT Press, Cambridge, Mass., 1994.

[10] G. Tian. The existence of equilibria in games with arbitrary strategy spaces and payoffs: a full characterization. Available at www.dklevine.com/archive/refs4814577000000000160.pdf., 2009.

How Many Levels Do Players Reason? Observational Challenges and a Solution[*]

[Extended Abstract] [*]

Adam Brandenburger
NYU Stern School of Business
New York University
adam.brandenburger@stern.nyu.edu

Alex Danieli
alex.daniels85@gmail.com

Amanda Friedenberg
Department of Economics
Arizona State University
amanda.friedenberg@asu.edu

Keywords

Epistemic Game Theory, Bounded Reasoning, Level-m Reasoning, Context of the Game

1. INTRODUCTION

Interactive reasoning is an important aspect of how players behave. To determine whether a particular course of action is good or bad, Ann may need to form a theory about Bob's play of the game. In forming such a theory, she may reason that Bob is 'strategically sophisticated' — if so, she may reason that Bob forms a belief about her own play to determine whether a particular strategy is good or bad for him. That is, Ann may want to form a second-order theory about Bob's play of the game. Of course, Ann may then reason that Bob uses a second-order theory to choose his strategy. In this case, Ann may want to form a third-order theory about Bob's play of the game. And so on.

How many levels of reasoning do players undertake? We address this question for the case where the players' processes of reasoning are not observable. Instead, the researcher observes only the behavior of the players — or, perhaps, only the outcome of the game. We ask: Can the researcher use this information to identify — or provide bounds on — levels of reasoning?

2. A MOTIVATING EXAMPLE

Figure 1 depicts the game of Battle-of-the-Sexes with an Outside Option: Ann can either choose to exercise an outside option or choose to play Battle-of-the-Sexes with Bob. The standard argument is that, if the players are 'strategically sophisticated,' then Ann will choose In-U and Bob will choose L: The strategy In-D is dominated for Ann by the outside option. Thus, if Ann does play In, Bob should reason that she will play U, since In-U is undominated. In this case, his best response is to play L. Ann should understand that Bob will reason this way and expect Bob to play L. With this expectation, Ann should play In-U.

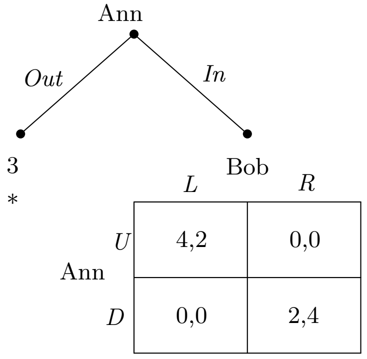


Figure 1: Battle-of-the-Sexes with an Outside Option

Thus, there appears to be a clear prediction: Ann will play In-U and Bob will play L. Yet, in the lab, Out is played with significant frequency. (See, inter alia, [4] and [2].) One might then draw the conclusion that there is limited reasoning: If Bob engages in one level of reasoning, he may choose either L or R depending on what he believes about Ann's play. So, if Ann engages in two levels of reasoning, she might choose to play Out. It is only if Ann reasons three (or more) levels that she would not play Out.

But, in fact, this behavior need not be an artifact of limited reasoning. There is another reason why Ann might choose to play Out — one based on the idea that there is a "context" to the game. In particular, suppose that it is commonly understood that "Bob is a bully" and, so, whenever a Battle-of-the-Sexes game is played, he attempts to go for

[*] This abstract was prepared for TARK 2015. We thank Jessy Hsieh, Terri Kneeland, Xiaomin Li, and Elliot Lipnowski for valuable comments. We also thank audiences at the University of Bristol, the 2013 Asian Meeting of the Econometric Society, National University of Singapore, and the 2013 Workshop on Logical Structure of Correlated Information Change for important input. Adam Brandenburger thanks the NYU Stern School of Business for financial support. Amanda Friedenberg thanks NSF SES-1358008 for financial support and the UCL Economics Department for its unbounded hospitality.

his best option and plays R. To be specific, suppose:

Bully-1: at each information set, Ann believes that Bob plays R,

Bully-2: at each information set, Bob believes "Bully-1,"

Bully-3: at each information set, Ann believes "Bully-2."

$\vdots$

If this is the context under which the game is played, Ann may play *Out*, even if she reasons three levels. In fact, she may play *Out*, even if she engages in common reasoning about the play of the game.

Here is the basic idea: When the game is played in this context, if Ann reasons at least one level, then she must play *Out*. (Ex ante, she expects that Bob will play R, and so *Out* is the unique best choice.) If Bob reasons at least two levels, then he will reason that Ann reasons at least one level, provided he has not observed information that contradicts this hypothesis. Thus, if Bob reasons at least two levels, he must begin the game believing Ann plays *Out*. Conditional upon Ann's playing *In*, he is forced to reason that Ann does not reason one level. So, Bob can reason two levels and conclude (after observing *In*) that Ann is playing *In*-U, but he can also conclude that Ann is playing *In*-D. Thus, if Bob reasons two levels, he can play either of L or R. It follows that believing that Bob play will R is consistent with three levels of reasoning for Ann.

In light of the above, it appears premature to conclude that observing the play of *Out* indicates that Ann is necessarily only a level m-reasoner, for some $m \leq 2$. This paper focuses on the case where the researcher cannot observe players' actual beliefs or which beliefs players consider possible. Thus, we seek an answer to our question that is independent of the context of the game.

3. APPROACH AND CHALLENGE

In keeping with what we have just seen, we will need to describe what beliefs players do vs. do not consider possible in a particular game. The device we will use to describe these beliefs is an **epistemic type structure**, denoted by $\mathcal{T}$. An epistemic type structure will consist of a set of types for each player, where a type for a player will describe that player's hierarchies of beliefs about the play of the game. Different type structures are associated with different events which are commonly believed. For instance, for Battle-of-the-Sexes with an Outside Option, there is a type structure where the event "Bob is a bully" is commonly believed and there are other type structures where it is not.

Call a pair $(\Gamma, \mathcal{T})$ an **epistemic game**. For a given epistemic game $(\Gamma, \mathcal{T})$, we can define the set of strategy-type pairs which are consistent with m levels of reasoning, to be denoted $R^m(\mathcal{T})$. Refer to Figure 2, that illustrates these sets. Specifically:

(0) The set $R^0(\mathcal{T})$ is the set of all strategy-type pairs. This set captures **level-**0 **reasoning**, since there is no requirement on reasoning.

(1) The set $R^1(\mathcal{T})$ is the set of strategy-type pairs where the players' strategies are optimal given their belief (i.e., type). This set captures **level-**1 **reasoning**.

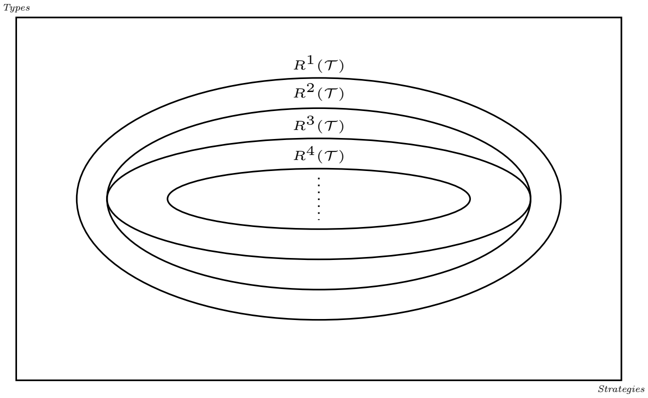


Figure 2: Level-m Reasoning

$\vdots$

(m) The set $R^{m+1}(\mathcal{T})$ is the set of strategy-type pairs in $R^m(\mathcal{T})$ where each player reasons that the other players engage in **level-**m **reasoning**. This set captures **level-**$(m+1)$ **reasoning**.

$\vdots$

(∞) The set $R^\infty(\mathcal{T})$ is the set of strategy-type pairs in $R^m(\mathcal{T})$ for all m. This set captures **level-**∞ or **common reasoning**.

The sets $R^m(\mathcal{T})$ depend not only on the game Γ but also on the type structure $\mathcal{T}$. This fits with our informal analysis of Battle-of-the-Sexes with an Outside Option, where the behavior of a level-3 reasoner depended on whether or not the event "Bob is a bully" is commonly understood.

Observe that the sets $R^0(\mathcal{T}), \ldots, R^m(\mathcal{T}), \ldots$ are decreasing. This reflects the fact that, if players reason at least $(m+1)$ levels, then they reason at least m levels. As a consequence, we will not be able to identify the minimum number of levels of reasoning by observing behavior alone.[1] Instead, we seek to identify the maximum number of levels of reasoning consistent with observed behavior.

The goal then is to identify when a strategy is consistent with m but not $(m+1)$ levels of reasoning. Toward this end, we seek to construct an ordered partition of the strategy set, denoted by $\mathcal{L} = \{L^0, L^1, \ldots, L^m, \ldots, L^\infty\}$, that satisfies the following criteria: When m is finite, $s \in L^m$ if it is

1. consistent with level-m reasoning in some epistemic game $(\Gamma, \mathcal{T}^*)$, but

2. inconsistent with level-$(m+1)$ reasoning in any epistemic game $(\Gamma, \mathcal{T})$.

When m is infinite, $s \in L^m$ if it is consistent with level-∞ reasoning in some epistemic game $(\Gamma, \mathcal{T}^*)$. We refer to players who choose $s \in L^m$ as **Level-m Reasoners** (or $\mathbf{L^m}$**-Reasoners**).

[1] As the "level-k" and "cognitive hierarchy" literatures make clear, it may be possible to identify the minimum number of levels of reasoning by making auxiliary assumptions about behavior or beliefs. See the discussion below.

For the case of a matrix, standard results give that $s \in L^m$ if and only if s is m but not $(m+1)$-rationalizable. One might conjecture that the same is true for the tree, where we now take "rationalizability" to mean "extensive-form rationalizability" as in [8] and [1] (or, equivalently, "iterated conditional dominance," as in [9]). This is not the case. Refer back to Battle-of-the-Sexes with an Outside Option (Figure 1). There, *Out* is consistent with two but not three rounds of extensive-form rationalizability. But, the "Bob is a bully" analysis showed that $Out \in L^\infty$.

The main paper provides a novel procedure which serves to construct the partition $\mathcal{L}$ in a finite number of steps. A major challenge in providing such an procedure is the fact that the definition of L^m makes reference to all type structures. For a given finite tree, there are infinitely many associated type structures and, therefore, searching across all type structures would appear to be an infinite task. The Main Theorem in the paper provides a way to side-step this difficulty for generic games. It characterizes the set of strategies consistent with m levels of reasoning as a property of the game alone — without reference to any type structure. It goes on to show how to implement the procedure in a "simple" manner.

4. UPPER BOUND ON REASONING

The main paper identifies the maximum number of levels of reasoning consistent with observed behavior. The focus on this upper bound is inevitable, absent making auxiliary assumptions on behavior and/or beliefs. (The "level-k" and "cognitive hierarchy" literatures, e.g., [7], [10], [5], [3], etc., obtain exact identification by imposing such restrictions.) But, there are also good reasons why this upper bound is of interest. To the extent that the researcher is interested in using the number of levels of reasoning as an empirical input, the researcher may care only that the player acts 'as if' she is an L^2-Reasoner — even if she is, in fact, an L^1-Reasoner or L^0-Reasoner. A similar argument applies to higher levels of reasoning.

For certain datasets, it may be possible to distinguish an L^2-Reasoner from an L^1-Reasoner who acts as if she is an L^2-Reasoner. With a large dataset, we should expect, on statistical grounds, to see more than just occasional in-sample play of strategies inconsistent with higher levels of reasoning. In the spirit of the "level-k" and "cognitive hierarchy" literatures (e.g., [7], [10], [5], [3], etc.), the researcher may be able to design an experiment so that, even in small samples, we would expect to see play inconsistent with higher levels of reasoning. For instance, the design in [6] is based on "ring games," where the payoff to any player depends only on his/her left-hand neighbor's choice. So, changing payoffs to a player two steps to the left of Ann should not affect her behavior if she is a L^1-Reasoner, but should affect her behavior is she is a L^2-Reasoner. As a consequence, varying payoffs across the experimental session should generate distinct behavior for L^1- and L^2-Reasoners, thereby allowing the experimenter to identify the L^2-Reasoner.

5. REFERENCES

[1] P. Battigalli. On Rationalizability in Extensive Games. *Journal of Economic Theory*, 74(1):40–61, 1997.

[2] J. Brandts and C. Holt. Limitations of dominance and forward induction: Experimental evidence. *Economics Letters*, 49(4):391–395, 1995.

[3] C. Camerer, T. Ho, and J. Chong. A Cognitive Hierarchy Model of Games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

[4] R. Cooper, D. DeJong, R. Forsythe, and T. Ross. Forward induction in the battle-of-the-sexes games. *American Economic Review*, 83(5):1303–1316, 1993.

[5] M. Costa-Gomes, V. Crawford, and B. Broseta. Cognition and Behavior in Normal-Form Games: An Experimental Study. *Econometrica*, 69(5):1193–1235, 2001.

[6] T. Kneeland. Testing Behavioral Game Theory: Higher-Order Rationality and Consistent Beliefs. http://terri.microeconomics.ca, 2013.

[7] R. Nagel. Unraveling in Guessing Games: An Experimental Study. *The American Economic Review*, 85(5):1313–1326, 1995.

[8] D. Pearce. Rationalizable Strategic Behavior and the Problem of Perfection. *Econometrica*, 52(4):1029–1050, 1984.

[9] M. Shimoji and J. Watson. Conditional Dominance, Rationalizability, and Game Forms. *Journal of Economic Theory*, 83(2):161–195, 1998.

[10] D. Stahl and P. Wilson. On Players' Models of Other Players: Theory and Experimental Evidence. *Games and Economic Behavior*, 10(1):218–254, 1995.

Translucent Players: Explaining Cooperative Behavior in Social Dilemmas

Valerio Capraro
Centre for Mathematics and Computer Science
(CWI)
Amsterdam, 1098 XG, The Netherlands
V.Capraro@cwi.nl

Joseph Y. Halpern
Cornell University
Computer Science Department
Ithaca, NY 14853
halpern@cs.cornell.edu

ABSTRACT

In the last few decades, numerous experiments have shown that humans do not always behave so as to maximize their material payoff. Cooperative behavior when non-cooperation is a dominant strategy (with respect to the material payoffs) is particularly puzzling. Here we propose a novel approach to explain cooperation, assuming what Halpern and Pass [2013] call *translucent players*. Typically, players are assumed to be *opaque*, in the sense that a deviation by one player in a normal-form game does not affect the strategies used by other players. But a player may believe that if he switches from one strategy to another, the fact that he chooses to switch may be visible to the other players. For example, if he chooses to defect in Prisoner's Dilemma, the other player may sense his guilt. We show that by assuming translucent players, we can recover many of the regularities observed in human behavior in well-studied games such as Prisoner's Dilemma, Traveler's Dilemma, Bertrand Competition, and the Public Goods game.

1. INTRODUCTION

In the last few decades, numerous experiments have shown that humans do not always behave so as to maximize their material payoff. Many alternative models have consequently been proposed to explain deviations from the money-maximization paradigm. Some of them assume that players are boundedly rational and/or make mistakes in the computation of the expected utility of a strategy [Camerer et al. 2004, Costa-Gomes et al. 2001, Halpern and Pass 2015, McKelvey and Palfrey 1995, Stahl and Wilson 1994]; yet others assume that players have other-regarding preferences [Bolton and Ockenfels 2000, Charness and Rabin 2002, Fehr and Schmidt 1999]; others define radically different solution concepts, assuming that players do not try to maximize their payoff, but rather try to minimize their regret [Halpern and Pass 2012, Renou and Schlag 2010], or maximize the forecasts associated to coalition structures [Capraro 2013, Capraro et al. 2013], or maximize the total welfare [Apt and Schäfer 2014, Rong and Halpern 2013]. (These references only scratch the surface; a complete bibliography would be longer than this paper!)

Cooperative behavior in one-shot anonymous games is particularly puzzling, especially in games where non-cooperation is a dominant strategy (with respect to the material payoffs): why should you pay a cost to help a stranger, when no clear direct or indirect reward seems to be at stake? Nevertheless, the secret of success of our societies is largely due to our ability to cooperate. We do not cooperate only with family members, friends, and co-workers. A great deal of cooperation can be observed also in one-shot anonymous interactions [Camerer 2003], where none of the five rules of cooperation proposed by Nowak [2006] seems to be at play.

Here we propose a novel approach to explain cooperation, based on work of Halpern and Pass [2013] and Salcedo [2013], assuming what Halpern and Pass call *translucent players*. Typically, players are assumed to be *opaque*, in the sense that a deviation by one player in a normal-form game does not affect the strategies used by other players. But a player may believe that if he switches from one strategy to another, the fact that he chooses to switch may be visible to the other players. For example, if he chooses to defect in Prisoner's Dilemma, the other player may sense his guilt. (Indeed, it is well known that there are facial and bodily clues, such as increased pupil size, associated with deception; see, e.g., [Ekman and Friesen 1969]. Professional poker players are also very sensitive to *tells*—betting patterns and physical demeanor that reveal something about a player's hand and strategy.)[1]

We use the idea of translucency to explain cooperation. This may at first seem somewhat strange. Typical lab experiments of social dilemmas consider anonymous players, who play each other over computers. In this setting, there are no tells. However, as Rand and his colleagues have argued (see, e.g., [Rand et al. 2012, Rand et al. 2014]), behavior of subjects in lab experiments is strongly influenced by their experience in everyday interactions. People internalize strategies that are more successful in everyday interactions and use them as default strategies in the lab. We would argue that people do not just internalize strategies; they also internalize *beliefs*. In everyday interactions, changing strategies certainly affects how other players react in the future. Through tells and possible leaks about changes in plans, it also may affect how other players react in current play. Thus, we would argue that, in everyday interactions, people assume a certain amount of transparency, both because it is a way of taking the future into account in real-world situations that are repeated and because it is a realistic assumption in one-shot games that are played in settings where

[1] The idea of translucency is motivated by some of the same concerns as Solan and Yariv's [2004] *games with espionage*, but the technical details are quite different. A game with espionage is a two-player extensive-form game that extends an underlying normal-form game by adding a step where player 1 can purchase some noisy information about player 2's planned move. Here, the information is free and all players may be translucent. Moreover, the effect of the translucency is modeled by the players' counterfactual beliefs rather than by adding a move to the game.

players have a great deal of social interaction. We claim that players then apply these beliefs in lab settings where they are arguably inappropriate.

There is experimental evidence that can be viewed as providing support for players believing that they are transparent. Gilovich et al. [1998] show that people tend to overestimate the extent to which others can discern their internal states. For instance, they showed that liars overestimate the detectability of their lies and that people believe that their feelings of disgust are more apparent than they actually are. There is also growing evidence that showing people simple images of watching eyes has a marked effect on behavior, ranging from giving more in Public Goods games to littering less (see [Bateson et al. 2013] for a discussion of some of this work and an extensive list of references). One way of understanding these results is that the eyes are making people feel more transparent.

We apply the idea of translucency to a particular class of games that we call *social dilemmas* (cf. [Dawes 1980]). A social dilemma is a normal-form game with two properties:

1. there is a unique Nash equilibrium s^N, which is a pure strategy profile;

2. there is a unique welfare-maximizing profile s^W, again a pure strategy profile, such that each player's utility if s^W is played is higher than his utility if s^N is played.

These uniqueness assumptions are not necessary, but they make definitions and computations easier. Although these restrictions are nontrivial, many of the best-studied games in the game-theory literature satisfy them, including Prisoner's Dilemma, Traveler's Dilemma [Basu 1994], Bertrand Competition, and the Public Goods game. (See Section 3 for more discussion of these games.)

There are (at least) two reasons why an agent may be concerned about translucency in a social dilemma: (1) his opponents may discover that he is planning to defect and punish him by defecting as well, (2) many other people in his social group (which may or may not include his opponent) may discover that he is planning to defect (or has defected, despite the fact that the game is anonymous) and think worse of him.

For definiteness, we focus here on the first point and assume that, in social dilemmas, players have a degree of belief α that they are transparent, so that if they intend to cooperate (by playing their component of the welfare-maximizing strategy) and decide to deviate, there is a probability α that another player will detect this, and play her component of the Nash equilibrium strategy. (The assumption that cooperation acts as a default strategy is supported by experiments showing that people forced to make a decision under time pressure are, on average, more cooperative than those forced to made a decision under time delay [Rand et al. 2012, Rand et al. 2014]. Rand and his colleagues suggest that this is due to the internalization of strategies that are successful in everyday interactions.) We assume that these detections are independent, so that the probability of, for example, exactly two players other than i detecting a deviation by i is $\alpha^2(1-\alpha)^{N-3}$, where N is the total number of players. Of course, if $\alpha = 0$, then we are back at the standard game-theoretic framework. We show that, with this assumption, we can already explain a number of experimental regularities observed in social dilemmas (see Section 3). We can model the second point regarding concerns about transparency in much the same way, and would get qualitatively similar results (see Section 6).

The rest of the paper is as follows. In Section 2, we formalize the notion of translucency in a game-theoretic setting. In Section 3, we define the social dilemmas that we focus on in this paper; in Section 4, we show that by assuming translucency, we can obtain as predictions of the framework a number of regularities that have been observed in the experimental literature. We discuss related work in Section 5. Section 6 concludes. Proofs are deferred to the full paper, where we also discuss a solution concept that we call *translucent equilibrium*, based on translucency, closely related to the notion of *individual rationality* discussed by Halpern and Pass [2013], and show how it can be applied in social dilemmas.

2. RATIONALITY WITH TRANSLUCENT PLAYERS

In this section, we briefly define rationality in the presence of translucency, motivated by the ideas in Halpern and Pass [2013].

Formally, a (finite) normal-form game $\mathcal{G}$ is a tuple $(P, S_1, \dots, S_N, u_1, \dots, u_N)$, where $P = \{1, \dots, N\}$ is the set of players, S_i is the set of strategies for player i, and u_i is player i's utility function. Let $S = S_1 \times \cdots \times S_N$ and $S_{-i} = \prod_{j \neq i} S_j$. We assume that S is finite and that $N \geq 2$.

In standard game theory, it is assumed that a player i has beliefs about the strategies being used by other players; i is rational if his strategy is a best response to these beliefs. The standard definition of best response is the following.

Definition 2.1. *A strategy $s_i \in S_i$ is a best response to a probability μ_i on S_{-i} if, for all strategies s_i' for player i,*

$$\sum_{s'_{-i} \in S_{-i}} \mu_i(s'_{-i}) u_i(s_i, s'_{-i}) \geq \sum_{s'_{-i} \in S_{-i}} \mu_i(s'_{-i}) u_i(s'_i, s'_{-i}).$$

□

Definition 2.1 implicitly assumes that i's beliefs about what other agents are doing do not change if i switches from s_i, the strategy he was intending to play, to a different strategy. (In general, we assume that i always has an *intended strategy*, for otherwise it does not make sense to talk about i switching to a different strategy.) So what we really have are beliefs $\mu_i^{s_i,s_i'}$ for i indexed by a pair of strategies s_i and s_i'; we interpret $\mu_i^{s_i,s_i'}$ as i's beliefs if he intends to play s_i but instead deviates to s_i'. Thus, $\mu_i^{s_i,s_i}$ represents i's beliefs if he plays s_i and does not deviate. We modify the standard definition of best response by defining best response with respect to a family of beliefs $\mu_i^{s_i,s_i'}$.

Definition 2.2. *Strategy $s_i \in S_i$ is a* best response for i with respect to the beliefs $\{\mu_i^{s_i,s_i'} : s_i' \in S_i\}$ *if, for all strategies $s_i' \in S_i$,*

$$\sum_{s'_{-i} \in S_{-i}} \mu_i^{s_i,s_i}(s'_{-i}) u_i(s_i, s'_{-i}) \geq \sum_{s'_{-i} \in S_{-i}} \mu_i^{s_i,s_i'}(s'_{-i}) u_i(s'_i, s'_{-i}).$$

□

We are interested in players who are making best responses to their beliefs, but we define best response in terms of Definition 2.2, not Definition 2.1. Of course, the standard notion of best response is just the special case of the notion above where $\mu_i^{s_i,s_i'} = \mu_i^{s_i,s_i}$ for all s_i': a player's beliefs about what other players are doing does not change if he switches strategies.

Definition 2.3. *A player is* translucently rational *if he best responds to his beliefs in the sense of Definition 2.2.* □

Our assumptions about translucency will be used to determine $\mu_i^{s_i,s_i'}$. For example, suppose that Γ is a 2-player game, player 1 believes that, if he were to switch from s_i to s_i', this would be detected by player 2 with probability α, and if player 2 did detect the switch, then player 2 would switch to s_j'. Then $\mu_i^{s_i,s_i'}$ is $(1-\alpha)\mu^{s_i,s_i} + \alpha\mu'$, where μ' assigns probability 1 to s_j'; that is, player 1 believes that with probability $1-\alpha$, player 2 continues to do what he would have done all along (as described by μ^{s_i,s_i}) and, with probability α, player 2 switches to s_j'.

3. SOCIAL DILEMMAS

Social dilemmas are situations in which there is a tension between the collective interest and individual interests: every individual has an incentive to deviate from the common good and act selfishly, but if everyone deviates, then they are all worse off. Many personal and professional relationships, the depletion of natural resources, climate protection, the security of energy supply, and price competition in markets can all be viewed as instances of social dilemmas.

As we said in the introduction, we formally define a social dilemma as a normal-form game with a unique Nash equilibrium and a unique welfare-maximizing profile, both pure strategy profiles, such that each player's utility if s^W is played is higher than his utility if s^N is played. While this is a quite restricted set of games, it includes many that have been quite well studied. Here, we focus on the following games:

Prisoner's Dilemma. Two players can either cooperate (C) or defect (D). To relate our results to experimental results on Prisoner's Dilemma, we think of cooperation as meaning that a player pays a cost $c > 0$ to give a benefit $b > c$ to the other player. If a player defects, he pays nothing and gives nothing. Thus, the payoff of (D, D) is $(0, 0)$, the payoff of (C, C) is $(b-c, b-c)$, and the payoffs of (D, C) and (C, D) are $(b, -c)$ and $(-c, b)$, respectively. The condition $b > c$ implies that (D, D) is the unique Nash equilibrium and (C, C) is the unique welfare-maximizing profile.

Traveler's Dilemma. Two travelers have identical luggage, which is damaged (in an identical way) by an airline. The airline offers to recompense them for their luggage. They may ask for any dollar amount between L and H (where L and H are both positive integers). There is only one catch. If they ask for the same amount, then that is what they will both receive. However, if they ask for different amounts—say one asks for m and the other for m', with $m < m'$—then whoever asks for m (the lower amount) will get $m+b$ (m and a bonus of b), while the other player gets $m-b$: the lower amount and a penalty of b. It is easy to see that (L, L) is the unique Nash equilibrium, while (H, H) maximizes social welfare, independent of b.

Public Goods game. $N \geq 2$ contributors are endowed with 1 dollar each; they must simultaneously decide how much, if anything, to contribute to a public pool. (The contributions must be in whole cent amounts.) The total contribution pot is then multiplied by a constant strictly between 1 and N, and then evenly redistributed among all players.[2] So the payoff of player i is $u_i(x_1, \ldots, x_N) = 1 - x_i + \rho(x_1 + \ldots + x_N)$, where x_i denotes i's contribution, and $\rho \in \left(\frac{1}{N}, 1\right)$ is the *marginal return*. (Thus, the pool is multiplied by ρN before being split evenly among all players.) Everyone contributing nothing to the pool is the unique Nash equilibrium, and everyone contributing their whole endowment to the pool is the unique welfare-maximizing profile.

Bertrand Competition. $N \geq 2$ firms compete to sell their identical product at a price between the "price floor" $L \geq 2$ and the "reservation value" H. (Again, we assume that H and L are integers, and all prices must be integers.) The firm that chooses the lowest price, say s, sells the product at that price, getting a payoff of s, while all other firms get a payoff of 0. If there are ties, then the sales are split equally among all firms that choose the lowest price. Now everyone choosing L is the unique Nash equilibrium, and everyone choosing H is the unique welfare-maximizing profile.[3]

From here on, we say that a player *cooperates* if he plays his part of the welfare-maximizing strategy profile and *defects* if he plays his part of the Nash equilibrium strategy profile.

While Nash equilibrium predicts that people should always defect in social dilemmas, in practice, we see a great deal of cooperative behavior; that is, people often play (their part of) the welfare-maximizing profile rather than (their part of) the Nash equilibrium profile. Of course, there have been many attempts to explain this. Evolutionary theories may explain cooperative behavior among genetically related individuals [Hamilton 1964] or when future interactions among the same subjects are likely [Nowak and Sigmund 1998, Trivers 1971]; see [Nowak 2006] for a review of the five rules of cooperation. However, we often observe cooperation even in one-shot anonymous experiments among unrelated players [Rapoport 1965].

Although we do see a great deal of cooperation in these games, we do not always see it. Here are some of the regularities that have been observed:

- The degree of cooperation in the Prisoner's dilemma depends positively on the benefit of mutual cooperation and negatively on the cost of cooperation [Capraro et al. 2014, Engel and Zhurakhovska 2012, Rapoport 1965].

[2] We thus consider only *linear* Public Goods games. This choice is motivated by the fact that our purpose is to compare the predictions of our model with experimental data. Most experiments have adopted linear Public Goods games, since they have much easier instructions and thus they minimize noise due to participants not understanding the rules of the game.

[3] We require that $L \geq 2$ for otherwise we would not have a unique Nash equilibrium, a condition we imposed on Social Dilemmas. If $L = 1$ and $N = 2$, we get two Nash equilibria: $(2, 2)$ and $(1, 1)$; similarly, for $L = 0$, we also get multiple Nash equilibria, for all values of $N \geq 2$.

- The degree of cooperation in the Traveler's Dilemma depends negatively on the bonus/penalty [Capra et al. 1999].

- The degree of cooperation in the Public Goods game depends positively on the constant marginal return [Gunnthorsdottir et al. 2007, Isaac et al. 1984].

- The degree of cooperation in the Public Goods game depends positively on the number of players [Barcelo and Capraro 2015, Isaac et al. 1994, Zelmer 2003].

- The degree of cooperation in the Bertrand Competition depends negatively on the number of players [Dufwenberg and Gneezy 2002].

- The degree of cooperation in the Bertrand Competition depends negatively on the price floor [Dufwenberg et al. 2007].

4. EXPLAINING SOCIAL DILEMMAS USING TRANSLUCENCY

As we suggested in the introduction, we hope to use translucency to explain cooperation in social dilemmas even when players cannot see each other. We expect that people get so used to assuming some degree of transparency in their everyday interactions, which are typically face-to-face, that they bring these strategies and beliefs in the lab setting, even though they are arguably inappropriate.

To do this, we have to make assumptions about an agent's beliefs. Say that an agent i has *type* (α, β, C) if i intends to cooperate (the parameter C stands for *cooperate*) and believes that (a) if he deviates from that, then each other agent will independently realize this with probability α; (b) if an agent j realizes that i is not going to cooperate, then j will defect; and (c) all other players will either cooperate or defect, and they will cooperate with probability β.

The standard assumption, of course, is that $\alpha = 0$. Our results are only of interest if $\alpha > 0$. The assumption that i believes that agent j will defect if she realizes that i is going to deviate from cooperation seems reasonable; defection is the "safe" strategy. We stress that, for our results, it does not matter what j actually does. All that matters are i's beliefs about what j will do. The assumption that players will either cooperate or defect is trivially true in Prisoner's Dilemma, but is a highly nontrivial assumption in the other games we consider. While cooperation and defection are arguably the most salient strategies, we do in practice see players using other strategies. For instance, the distribution of strategies in the Public Goods game is typically trimodal, concentrated on contributing nothing, contributing everything, and contributing half [Capraro et al. 2014]. We made this assumption mainly for technical convenience: it makes the calculations much easier. We believe that results qualitatively similar to ours will hold under a much weaker assumption, namely, that a type (α, β, C) player believes that other players will cooperate with probability β (without assuming that they will defect with probability $1 - \beta$).

Similarly, the assumptions that a social dilemma has a unique Nash equilibrium and a unique social-welfare maximizing strategy were made largely for technical reasons. We can drop these assumptions, although that would require more complicated assumptions about players' beliefs.

Our assumptions ensure that the type of player i determines the distributions $\mu_i^{s_i,s_i'}$. In a social dilemma with N agents, the distribution $\mu_i^{s_i,s_i}$ assigns probability $\beta^r(1-\beta)^{N-1-r}$ to a strategy profile s_{-i} for the players other than i if exactly r players cooperate in s_{-i} and the remaining $N-1-r$ players defect; it assigns probability 0 to all other strategy profiles. The distributions $\mu_i^{s_i,s_i'}$ for $s_i' \neq s_i$ all have the form $\sum_{J \subseteq \{1,\ldots,i-1,i+1,\ldots,N\}} \alpha^{|J|}(1-\alpha)^{N-1-|J|}\mu_i^J$, where μ_i^J is the distribution that assigns probability $\beta^k(1-\beta)^{N-|J|-k}$ to a profile where $k \leq N-1-|J|$ players not in J cooperate, and the remaining players (which includes all the players in J) defect. Thus, μ_i^J is the distribution that describes what player i's beliefs would be if he knew that exactly the players in J had noticed his deviation (which happens with probability $\alpha^{|J|}(1-\alpha)^{N-1-|J|}$). In the remainder of this section, when we talk about best response, it is with respect to these beliefs.

For our purposes, it does not matter where the beliefs α and β that make up a player's type come from. We do not assume, for example, that other players are (translucently) rational. For example, i may believe that some players cooperate because they are altruistic, while others may cooperate because they have mistaken beliefs. We can think of β as summarizing i's previous experience of cooperation when playing social dilemmas. Here we are interested in the impact of the parameters of the game on the reasonableness of cooperation, given a player's type.

The following four propositions analyze the four social dilemmas in turn; the proofs can be found in the full paper. We start with Prisoner's Dilemma. Recall that b is the benefit of cooperation and c is its cost.

PROPOSITION 4.1. *In Prisoner's Dilemma, it is translucently rational for a player of type (α, β, C) to cooperate if and only if $\alpha\beta b \geq c$.* □

As we would expect, if $\alpha = 0$, then cooperation is not a best response in Prisoner's Dilemma; this is just the standard argument that defection dominates cooperation. But if $\alpha > 0$, then cooperation can be rational. Moreover, if we fix α, the greater the benefit of cooperation and the smaller the cost, then the smaller the value of β that still allows cooperation to be a best response.

We next consider Traveler's Dilemma. Recall that b is the reward/punishment, H is the high payoff, and L is the low payoff,

PROPOSITION 4.2. *In Traveler's Dilemma, it is translucently rational for a player of (α, β, C) to cooperate if and only if*

$$b \leq \begin{cases} \frac{(H-L)\beta}{1-\alpha\beta} & \text{if } \alpha \geq \frac{1}{2} \\ \min\left(\frac{(H-L)\beta}{1-\alpha\beta}, \frac{H-L-1}{1-2\alpha}\right) & \text{if } \alpha < \frac{1}{2}. \end{cases}$$

□

Proposition 4.2 shows that as b, the punishment/reward, increases, a player must have greater belief that his opponent is cooperative and/or a greater belief that the opponent will learn about his deviation and/or a greater difference between the high and low payoffs in order to make cooperation a best response. (The fact that increasing β increases $\frac{(H-L)\beta}{1-\alpha\beta}$ follows from straightforward calculus.)

We next consider the Public Goods game. Recall the ρ is the marginal return of cooperating.

Proposition 4.3. *In the Public Goods game with N players, it is translucently rational for a player of type (α, β, C) to cooperate if and only if $\alpha\beta\rho(N-1) \geq 1-\rho$.* □

Proposition 4.3 shows that if $\rho = 1$, then cooperation is certainly a best response (you always get out at least as much as you contribute). For fixed α and β, there is guaranteed to be an N_0 such that cooperation is a best response for all $N \geq N_0$; moreover, for fixed α, as N gets larger, smaller and smaller βs are needed for cooperation to be a best response.

Finally we consider the Bertrand competition. Recall that H is the reservation value and L is the price floor.

Proposition 4.4. *In Bertrand Competition, it is translucently rational for a player of type (α, β, C) to cooperate iff $\beta^{N-1} \geq \max(\gamma^{N-1}N(H-1)/H, f(\gamma, N)LN/H)$, where $\gamma = (1-\alpha)\beta$ and $f(\gamma, N) = \sum_{k=0}^{N-1} \binom{N-1}{k}(1-\gamma)^k\gamma^{N-k-1}/(k+1)$.* □

Note that $f(\gamma, N) = \sum_{k=0}^{N-1} \binom{N-1}{k}(1-\gamma)^k\gamma^{N-k-1}/(k+1) \geq \sum_{k=0}^{N-1} \binom{N-1}{k}(1-\gamma)^k\gamma^{N-k}/N = 1/N$, so Proposition 4.4 shows cooperation is irrational if $\beta^{N-1} < L/H$. Thus, while cooperation may be achieved for reasonable values of α and β if N is small, a player must be more and more certain of cooperation in order to cooperate in Bertrand Competition as the number of players increases. Indeed, for a fixed type (α, β, C), there exists N_0 such that cooperation is not a best response for all $N \geq N_0$. Moreover, if we fix the number N of players, more values of α and β allow cooperation as L/H gets smaller. In particular, if we fix H and raise the floor L, fewer values of α and β allow cooperation.

While Propositions 4.1–4.4 are suggestive, we need to make extra assumptions to use these propositions to make predictions. A simple assumption that suffices is that there are a substantial number of translucently rational players whose types have the form (α, β, C), and for each pair (u, v) and (u', v') of open intervals in $[0, 1]$, there is a positive probability of finding someone of type (α, β, C) with $\alpha \in (u, v)$ and $\beta \in (u', v')$. With this assumption, it is easy to see that all the regularities discussed in Section 3 hold.

5. COMPARISON TO OTHER APPROACHES

Here we show that approaches (that we are aware of) other than that of Charness and Rabin and possibly that of Bolton and Ockenfels are not able to obtain all the regularities that we mentioned in Section 3. We consider a number of approaches in turn.

- The Fehr and Schmidt [1999] *inequity-aversion model* assumes that subjects play a Nash equilibrium of a modified game, in which players do not only care about their monetary payoff, but also they care about equity. Specifically, player i's utility when strategy s is played is assumed to be $U_i(s) = u_i(s) - \frac{a_i^{FS}}{N-1}\sum_{j\neq i}\max(u_j(s)-u_i(s), 0) - \frac{b_i^{FS}}{N-1}\sum_{j\neq i}\max(u_i(s)-u_j(s), 0)$, where $u_i(s)$ is the material payoff of player i, and $0 \leq b_i^{FS} \leq a_i^{FS}$ are individual parameters, where a_i^{FS} represents the extent to which player i is averse to inequity in favor of others, and b_i^{FS} represents his aversion to inequity in his favor. Consider the Public Goods game with N players. The strategy profile $(x, \dots, x)$, where all players contribute x gives player i a utility of $(1-x)+\rho Nx$. If $x > 0$ and player i contributes $x' < x$, then his payoff is $(1-x') + \rho((N-1)x + x') - b_i^{FS}\rho(x-x')$. Thus, $(x, \dots, x)$ is an equilibrium if $b_i^{FS}\rho(x-x') \geq (1-\rho)(x-x')$, that is, if $b_i^{FS} \geq (1-\rho)/\rho$. Thus, if $b_i^{FS} \geq (1-\rho)/\rho$ for all players i, then $(x, \dots, x)$ is an equilibrium for all choices of x and all values of N. While there may be other pure and mixed strategy equilibria, it is not hard to show that if $b_i^{FS} < (1-\rho)/\rho$, then player i will play 0 in every equilibrium (i.e., not contribute anything). As a consequence, assuming, as in our model, that players believe that there is a probability β that other agents will cooperate and that the other agents either cooperate or defect, Fehr and Schmidt [1999] model does not make any clear prediction of a group-size effect on cooperation in the public goods game.

- McKelvey and Palfrey's [1995] *quantal response equilibrium (QRE)* is defined as follows.[4] Taking $\sigma_i(s)$ to be the probability that mixed strategy σ_i assigns to the pure strategy s, given $\lambda > 0$, a mixed strategy profile σ is a QRE if, for each player i, $\sigma_i(s) = \frac{e^{\lambda EU_i(s,\sigma_{-i})}}{\sum_{s'_i \in S_i} e^{\lambda EU_i(s'_i,\sigma_{-i})}}$.

  To see that QRE does not describe human behaviour well in social dilemmas, observe that in the Prisoner's Dilemma, for all choices of parameters b and c in the game, all choices of the parameter λ, all players i, and all (mixed) strategies s_{-i} of player $-i$, we have $EU_i(C, s_{-i}) < EU_i(D, s_{-i})$. Consequently, whatever the QRE σ is, we must have $\sigma_i(C) < \frac{1}{2} < \sigma_i(D)$, that is, QRE predicts that the degree of cooperation can never be larger than 50%. However, experiments show that we can increase the benefit-to-cost ratio so as to reach arbitrarily large degrees of cooperation (close to 80% in [Capraro et al. 2014] with $b/c = 10$).

- *Iterated regret minimization* [Halpern and Pass 2012] does not make appropriate predictions in Prisoner's Dilemma and the Public Goods game, because it predicts that if there is a dominant strategy then it will be played, and in these two games, playing the Nash equilibrium is the unique dominant strategy.

- Capraro's [2013] notion of *cooperative equilibrium*, while correctly predicting the effects of the size of the group on cooperation in the Bertrand Competition and the Public Goods game [Barcelo and Capraro 2015], fails to predict the negative effect of the price floor on cooperation in the Bertrand Competition.

- Rong and Halpern's [2010, 2013] notion of *cooperative equilibrium* (which is different from that of Capraro [2013]) focuses on 2-player games. However, the definition for games with greater than 2 players does not predict the decrease in cooperation as N increases in

[4] We actually define here a particular instance of QRE called the *logit QRE*; λ is a free parameter of this model.

Bertrand Competition, nor the increase as N increases in the Public Goods Game.

- Bolton and Ockenfels' [2000] *inequity-aversion model* assumes that a player i aims at maximizing his or her *motivational function* $v_i = v_i(x_i, \sigma_i)$, where x_i is i's monetary payoff and $\sigma_i = \sigma_i(x_1, \sum_{j=1,\ldots,N} x_j) = x_i / \sum_{j=1,\ldots,N} x_j$. The motivational function is assumed to be twice differentiable, weakly increasing in the first argument, and concave in the second argument with a maximum at $\sigma_i = \frac{1}{N}$, but otherwise is unconstrained. For each of the social dilemmas that we have considered, it is not hard to define a motivational function that will obtain the regularities observed. However, we have not been able to find a single motivational function that gives the observed regularities for all four social dilemmas that we have considered. In any case, just as with the Charness and Rabin model, once we consider the interaction between social groups and translucency, we can distinguish our approach from this inequity-aversion model. Specifically, consider a situation where people are given a choice between giving \$1 to an anonymous stranger, rather than burning it. In such a situation, inequity aversion would predict that people would burn the dollar to maintain equity (i.e., a situation where no one gets \$1). However, perhaps not surprisingly, Capraro et al. [2014] found that over 90% people prefer giving away the dollar to burning it. Of course, translucency (and a number of other approaches) would have no difficulty in explaining this phenomenon.

The one approach besides ours that we are aware of that obtains all the regularities discussed above is that of Charness and Rabin [2002]. Charness and Rabin, like Fehr and Schmidt [1999], assume that agents play a Nash equilibrium of a modified game, where players care not only about their personal material payoff, but also about the social welfare and the outcome of the least fortunate person. Specifically, player i's utility is assumed to be $(1 - a_i^{CR})u_i(s) + a_i^{CR}(b_i^{CR} \min_{j=1,\ldots,N} u_j(s) + (1 - b_i^{CR}) \sum_{j=1}^{N} u_j(s))$. Assuming, as in our model, that agents believe that other players either cooperate or defect and that they cooperate with probability β, then it is not hard to see that Charness and Rabin [2002] also predict all the regularities that we have been considering.

Although it seems difficult to distinguish our model from that of Charness and Rabin [2002] if we consider only social dilemmas, the models are distinguishable if we look at other settings and take into account the other reason we mentioned for translucency: that other people in their social group might discover how they acted. We can easily capture this in the framework we have been considering by doubling the number of agents; for each player i, we add another player i^* that represent's i's social network. Player i^* can play only two actions: n (for "did *n*ot observe player i's action) and o (for "*o*bserved player i's action").[5] The payoffs of these new players are irrelevant. Player i's payoff depends on the action of player i^*, but not on the actions of player j^* for $j^* \neq i^*$. Now player i must have a prior probability γ_i about whether his action will be observed; in a social dilemma, this probability might increase to $\gamma_i' \geq \gamma_i$ if he intends to cooperate but instead deviates and defects. It should be clear that, even if $\gamma_i' = \gamma_i$, if we assume that player i's utilities are significantly lower if his non-cooperative action is observed, with this framework we would get qualitatively similar results for social dilemmas to the ones that we have already obtained. Again, a player has beliefs about the extent to which he is transparent, and we can set the payoffs so that the effects of transparency are the same if a player's social network learns about his actions and if other players learn about his action.

The advantage of taking into account what your social group thinks is that it allows us to apply ideas of translucency even to single-player games like the Dictator Game [Kahneman et al. 1986]. To do so, we need to make assumptions about what a player's utility would be if his social group knew the extent to which he shared the pot. But it should be clear that reasonable assumptions here would lead to some degree of sharing. While this would still not distinguish our predictions from those of the Charness-Rabin model, there is a variant of the Dictator Game that has recently been considered to show existence of hyper-altruism in conflict situations [Crockett et al. 2014, Capraro 2014]. In the simplest version of this game, there are only two possible allocations of money: either the agent gets x and the other player gets $-x$, or the other player gets x and the agent gets $-x$. In this game, the Charness-Rabin approach would predict that the agent will either keep x or be indifferent between keeping x and giving it away. But assuming translucency allows for the possibility that some types of agents would think that their social group would approve of them giving away x, so if the action were observed by their social group, they would get high utility by giving away x. However, recent results by Capraro [2014] show that a significant fraction (1/6) of people are *hyper-altruistic*: they strictly prefer giving away x to keeping it [Capraro 2014].

Just to be clear, we do not mean to imply that translucency is the unique "right" explanation for cooperation in social dilemmas and all the other explanations that we discussed above are "wrong". There are probably a number of factors that contribute to cooperation. We hope in future work to tease these apart.

6. DISCUSSION

We have presented an approach that explains a number of well-known observations regarding the extent of cooperation in social dilemmas. In addition, our approach can also be applied to explain the apparent contradiction that people cooperate more in a one-shot Prisoner's dilemma when they do not know the other player's choice than when they do. In the latter case, Shafir and Tversky [1992] found that most people (90%) defect, while in the former case, only 63% of people defect. Our model of translucent players predicts this behavior: if player 1 knows player 2's choices then there is no translucency, so our model predicts that player 1 defects for sure. On the other hand, if player 1 does not know player 2's choice and believes that he is to some extent translucent, then, as shown in Proposition 4.1, he may be willing to cooperate. Seen in this light, our model can also be interpreted as an attempt to formalize *quasi-magical thinking* [Shafir and Tversky 1992], the kind of reasoning that is supposed to motivate those people who believe that the others' reasoning is somehow influenced by their own thinking, even

[5] Alternatively, we could take player i's payoff to depend on the state of the world, where the state would model whether or not player i's action was observed.

though they know that there is no causal relation between the two. Quasi-magical thinking has also been formalized by Masel [2007] in the context of the Public Goods game and by Daley and Sadowski [2014] in the context of symmetric 2×2 games. The notion of translucency goes beyond these models, since it may be applied to a much larger set of games.

Besides a retrospective explanation, our model makes new predictions for social dilemmas which, to the best of our knowledge, have never been tested in the lab. In particular, it predicts that

- the degree of cooperation in Traveler's dilemma increases as the difference $H - L$ increases;
- for fixed L and N, the degree of cooperation in Bertrand Competition increases as H increases, and what really matters is the ratio L/H.

Clearly much more experimental work needs to be done to validate the approach. For one thing, it is important to understand the predictions it makes for other social dilemmas and for games that are not social dilemmas. Perhaps even more important would be to see if we can experimentally verify that people believe that they are to some extent translucent, and, if so, to get a sense of what the value of α is. In light of the work on watching eyes mentioned in the introduction, it would also be interesting to know what could be done to manipulate the value of α.

One feature of our approach is that, at least if we take the concern with translucency to be due to an opponent discovering what you are going to do (rather than other members of your social group discovering what you are going to do), then, unlike many other approaches to explaining social dilemmas, our approach does not involve modifying the utility function; that is, we can apply translucency while still identifying utility with the material payoff. While this make it an arguably simpler explanation, that does not necessarily make it "right", of course. We do not in fact believe that there is a unique "right" explanation for cooperation in social dilemmas and all the other explanations that we discussed above are "wrong". There are probably a number of factors that contribute to cooperation. We hope in future work to tease these apart.

Of course, we do not have to assume $\alpha > 0$ to get cooperation in social dilemmas such as Traveler's Dilemma or Bertrand Competition. But we do if we want to consider what we believe is the appropriate equilibrium notion. Suppose that rational players are chosen at random from a population and play a social dilemma. Players will, of course, then update their beliefs about the likelihood of seeing cooperation, and perhaps change their strategy as a consequence. Will these beliefs stabilize and the strategies played stabilize? By *stability* here, we mean that (1) players are all best responding to their beliefs, and (2) players' beliefs about the strategies played by others are correct: if player i ascribes probability p to player j playing a strategy s_j, then in fact a proportion p of players in the population play s_j. We have deliberately been fuzzy here about whether we mean best response in the sense of Definition 2.1 or Definition 2.2. If we use Definition 2.1 (or, equivalently use Definition 2.2 and take $\alpha = 0$), then it is easy to see (and well known) that the only way that this can happen is if the distribution of strategies played by the players represents a mixed strategy Nash equilibrium. On the other hand, if $\alpha > 0$ and we use Definition 2.2, then we can have stable beliefs that accurately reflect the strategies used and have cooperation (in all the other social dilemmas that we have studied). We make this precise in the full paper, using the framework of Halpern and Pass [2013], by defining a notion of *translucent equilibrium.* Roughly speaking, we construct a model where, at all states, players are translucently rational (so we have common belief of translucent rationality), the strategies used are common knowledge, and we nevertheless have cooperation at some states. Propositions 4.1–4.4 play a key role in this construction; indeed, as long as the strategies used satisfy the constraints imposed by these results, we get a translucent equilibrium.

In the full paper, we also characterize those profiles of strategies that can be translucent equilibria, using ideas similar in spirit to those of Halpern and Pass [2013]. While allowing people to believe that they are to a certain extent transparent means that the set of translucent equilibria is a superset of the set of Nash equilibria, not all strategy profiles can be translucent equilibria. For example, (C,D) is not a translucent equilibrium in Prisoner's dilemma. We have not focused on translucent equilibrium in the main text, because it makes strong assumptions about players' rationality and beliefs (e.g., it implicitly assumes common belief of translucent rationality). We do not need such strong assumptions for our results.

7. REFERENCES

[2014] Apt, K. and G. Schäfer (2014). Selfishness level of strategic games. *Journal of Artificial Intelligence Research 49*, 207–240.

[2015] Barcelo, H. and V. Capraro (2015). Group size effect on cooperation in one-shot social dilemmas. *Scientific Reports 5.*

[1994] Basu, K. (1994). The traveler's dilemma: paradoxes of rationality in game theory. *American Economic Review 84*(2), 391–395.

[2013] Bateson, M., L. Callow, J. R. Holmes, M. L. Redmond Roche, and D. Nettle (2013). Do images of "watching eyes" induce behaviour that is more pro-social or more normative? A field experiment on littering. *PLoS ONE 8*(12).

[2000] Bolton, G. E. and A. Ockenfels (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review 90*(1), 166–193.

[2003] Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton, N. J.: Princeton University Press.

[2004] Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics 119*, 861–897.

[1999] Capra, M., J. K. Goeree, R. Gomez, and C. A. Holt (1999). Anomalous behavior in a traveler's dilemma. *American Economic Review 89*(3), 678–690.

[2013] Capraro, V. (2013). A model of human cooperation in social dilemmas. *PLoS ONE 8*(8), e72427.

[2014] Capraro, V. (2014). The emergence of altruistic behaviour in conflictual situations. Working Paper.

[2014] Capraro, V., J. J. Jordan, and D. G. Rand (2014). Heuristics guide the implementation of social preferences in one-shot Prisoner's Dilemma experiments. *Scientific*

Reports 4.

[2014] Capraro, V., C. Smyth, K. Mylona, and G. A. Niblo (2014). Benevolent characteristics promote cooperative behaviour among humans. *PloS ONE 9*(8), e102881.

[2013] Capraro, V., M. Venanzi, M. Polukarov, and N. R. Jennings (2013). Cooperative equilibria in iterated social dilemmas. In *Proc. Sixth International Symposium on Algorithmic Game Theory (SAGT '13)*, pp. 146–158.

[2002] Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics 117*(3), 817–869.

[2001] Costa-Gomes, M., V. Crawford, and B. Broseta (2001). Cognition and behavior in normal form games: An experimental study. *Econometrica 69*(5), 1193–1235.

[2014] Crockett, M. J., Z. Kurth-Nelson, J. Z. Siegel, P. Dayan, and R. J. Dolan (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences 111*(48), 17320–17325.

[2014] Daley, B. and P. Sadowski (2014). A strategic model of magical thinking: Axioms and analysis. Available at http://www.princeton.edu/economics/seminar-schedule-by-prog/behavioralf14/Daley_Sadowski_MT.pdf.

[1980] Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology 31*, 169–193.

[2002] Dufwenberg, M. and U. Gneezy (2002). Information disclosure in auctions: an experiment. *Journal of Economic Behavior and Organization 48*, 431–444.

[2007] Dufwenberg, M., U. Gneezy, J. K. Goeree, and R. Nagel (2007). Price floors and competition. *Special Issue of Economic Theory 33*, 211–224.

[1969] Ekman, P. and W. Friesen (1969). Nonverbal leakage and clues to deception. *Psychiatry 32*, 88–105.

[2012] Engel, C. and L. Zhurakhovska (2012). When is the risk of cooperation worth taking? The Prisoner's Dilemma as a game of multiple motives. Working Paper.

[1999] Fehr, E. and K. Schmidt (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics 114*(3), 817–868.

[1998] Gilovich, T., K. Savitsky, and V. H. Medvec (1998). The illusion of transparency: biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology 75*(2), 332.

[2007] Gunnthorsdottir, A., D. Houser, and K. McCabe (2007). Dispositions, history and contributions in public goods experiments. *Journal of Economic Behavior and Organization 62*(2), 304–315.

[2012] Halpern, J. Y. and R. Pass (2012). Iterated regret minimization: a new solution concept. *Games and Economic Behavior 74*(1), 194–207.

[2013] Halpern, J. Y. and R. Pass (2013). Game theory with translucent players. In *Theoretical Aspects of Rationality and Knowledge: Proc. Fourteenth Conference (TARK 2013)*, pp. 216–221.

[2015] Halpern, J. Y. and R. Pass (2015). Algorithmic rationality: Game theory with costly computation. *Journal of Economic Theory 156*, 246–268.

[2010] Halpern, J. Y. and N. Rong (2010). Cooperative equilibrium. In *Proc. Ninth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 1465–1466.

[1964] Hamilton, W. D. (1964). The genetical evolution of social behavior. i. *Journal of Theoretical Biology 7*, 1–16.

[1984] Isaac, M. R., J. M. Walker, and S. Thomas (1984). Divergent evidence on free riding: an experimental examination of possible explanations. *Public Choice 43*(1), 113–149.

[1994] Isaac, M. R., J. M. Walker, and A. W. Williams (1994). Group size and the voluntary provision of public goods. *Journal of Public Economics 54*, 1–36.

[1986] Kahneman, D., J. Knetsch, and R. H. Thaler (1986). Fairness and the assumptions of economics. *Journal of Business 59*(4), S285–300.

[2007] Masel, J. (2007). A Bayesian model of quasi-magical thinking can explain observed cooperation in the public good game. *Journal of Economic Behavior and Organization 64*(1), 216–231.

[1995] McKelvey, R. and T. Palfrey (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior 10*(1), 6–38.

[2006] Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science 314*(5805), 1560–1563.

[1998] Nowak, M. A. and K. Sigmund (1998). Evolution of indirect reciprocity by image scoring. *Nature 393*, 573–577.

[2012] Rand, D. G., J. D. Green, and M. A. Nowak (2012). Spontaneous giving and calculated greed. *Nature 489*, 427–430.

[2014] Rand, D. G., A. Peysakhovich, G. T. Kraft-Todd, G. E. Newman, O. Wurzbacher, M. A. Nowak, and J. D. Greene (2014). Social heuristics shape intuitive cooperation. *Nature Communications 5*, 3677.

[1965] Rapoport, A. (1965). *Prisoner's Dilemma: A Study in Conflict and Cooperation*. University of Michigan Press.

[2010] Renou, L. and K. H. Schlag (2010). Minimax regret and strategic uncertainty. *Journal of Economic Theory 145*, 264–286.

[2013] Rong, N. and J. Y. Halpern (2013). Towards a deeper understanding of cooperative equilibrium: characterization and complexity. In *Proc. Twelfth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 319–326.

[2013] Salcedo, B. (2013). Implementation without commitment in moral hazard environments. Working paper.

[1992] Shafir, E. and A. Tversky (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology 24*, 449–474.

[2004] Solan, E. and L. Yariv (2004). Games with espionage. *Games and Economic Behavior 47*, 172–199.

[1994] Stahl, D. and P. Wilson (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization 25*(3), 309–327.

[1971] Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology 46*, 35–57.

[2003] Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics 6*, 299–310.

Announcement as effort on topological spaces

Hans van Ditmarsch
hans.van-ditmarsch@loria.fr

Sophia Knight
sophia.knight@gmail.com

Aybüke Özgün
aybuke.ozgun@loria.fr

LORIA, CNRS - Université de Lorraine

ABSTRACT

We propose a multi-agent logic of knowledge, public and arbitrary announcements, that is interpreted on topological spaces in the style of subset space semantics. The arbitrary announcement modality functions similarly to the effort modality in subset space logics, however, it comes with intuitive and semantic differences. We provide axiomatizations for three logics based on this setting, and demonstrate their completeness.

Keywords

Topology, subset space logic, dynamic epistemic logic, arbitrary (public) announcements

1. INTRODUCTION

In [13], Moss et al. introduce a bi-modal logic with language

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K\varphi \mid \Box\varphi,$$

called subset space logic (SSL), in order to formalize reasoning about sets and points together in one modal system. The main interest in their investigation lies in spatial structures such as topological spaces and using modal logic and the techniques behind for spatial reasoning, however, they also have a strong motivation from epistemic logic. While the modality K is interpreted as knowledge, $\Box$ intends to capture the notion of *effort*, i.e., any action that results in increase in knowledge. They propose subset space semantics for their logic. A subset space is defined to be a pair $(X, \mathcal{O})$, where X is a non-empty domain and $\mathcal{O}$ is a collection of subsets of X (not necessarily a topology), wherein the modalities K and $\Box$ are evaluated with respect to pairs of the form (x, U), where $x \in U \in \mathcal{O}$. According to subset space semantics, given a pair (x, U), the modality K quantifies over the elements of U, whereas $\Box$ quantifies over all open subsets of U that include the actual world x. Therefore, while knowledge is interpreted 'locally' in a given observation set U, effort is read as *open-set-shrinking* where more effort corresponds to a smaller neighbourhood, thus, a possible increase in knowledge. The schema $\Diamond K\varphi$ states that after some effort the agent comes to know φ where effort can be in the form of measurement, observation, computation, approximation [13, 8, 14, 5], or announcement [15, 1, 16].

The epistemic motivation behind the subset space semantics and the dynamic nature of the effort modality suggests a link between SSL and dynamic epistemic logic, in particular dynamics known as public announcement [4, 5, 3, 19, 6]. The works [4, 5, 3] propose modelling public announcements on subset spaces by deleting the states or the neighbourhoods falsifying the announcement. This dynamic epistemic method is not in the spirit of the effort modality: dynamic epistemic actions result in global model change, whereas the effort modality results in local neighbourhood shrinking. Hence, it is natural to search for an 'open-set-shrinking-like' interpretation of public announcements on subset spaces. To best of our knowledge, Wang and Ågotnes [19] were the first to propose semantics for public announcements on subset spaces in the style of the effort modality, although this is not necessarily on topological spaces. Bjorndahl [6] then proposed a revised version of the [19] semantics. In contrast to the aforementioned proposals, Bjorndahl uses models based on topological spaces to interpret knowledge and information change via public announcements. He considers the language

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K\varphi \mid int(\varphi) \mid [\varphi]\varphi,$$

where $int(\varphi)$ means 'φ is true and can be announced', and where $[\varphi]\psi$ means 'after public announcement of φ, ψ.'

In [1], Balbiani et al. introduce a logic to quantify over announcements in the setting of epistemic logic based on the language (with single-agent version here)

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K\varphi \mid [\varphi]\varphi \mid \Box\varphi.$$

In this case, unlike above, $\Box\varphi$ means 'after any announcement, φ (is true)' so that $\Box$ quantifies over epistemically definable subsets ($\Box$-free formulas of the language) of a given model. In this case, $\Diamond K\varphi$ again means that the agent comes to know φ, but in the interpretation that there is a formula ψ such that after announcing it the agent knows φ. What becomes true or known by an agent after an announcement can be expressed in this language without explicit reference to the announced formula.

Clearly, the meaning of the effort $\Box$ modality and of the arbitrary announcement $\Box$ modality are related in motivation. In both cases, interpreting the modality requires quantification over sets. Subset-space-like semantics provides natural tools for this. In [16], we extended Bjorndahl's proposal [6] with an arbitrary announcement modality

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K\varphi \mid int(\varphi) \mid [\varphi]\varphi \mid \Box\varphi$$

and provided topological semantics for the $\Box$ modality, and proved completeness for the corresponding single-agent logic $APAL_{int}$.

In the current proposal we generalize this approach to a multi-agent setting. Multi-agent subset space logics have been investigated in [11, 12, 4, 18]. There are some challenges with such a logic concerning the evaluation of higher-order knowledge. The general setup is for any finite number of agents, but to demonstrate the challenges, consider the case of two agents. Suppose for each of two agents i and j there is an open set such that the semantic primitive becomes a triple (x, U_i, U_j) instead of a pair (x, U). Now consider a formula like $K_i\hat{K}_jK_ip$, for 'agent i knows that agent j considers possible that agent i knows proposition p'. If this is true for a triple (x, U_i, U_j), then $\hat{K}_jK_ip$ must be true for any $y \in U_i$; but y may not be in U_j, in which case (y, U_i, U_j) is not well-defined: we cannot interpret $\hat{K}_jK_ip$. Our solution to this dilemma is to consider neighbourhoods that are not only relative to each agent, as usual in multi-agent subset space logics, but that are also *relative to each state*. This amounts to, when shifting the viewpoint from x to $y \in U_i$, in (x, U_i, U_j), we simultaneously have to shift the *neighbourhood* (and not merely the point in the actual neighbourhood) for the other agent. So we then go from (x, U_i, U_j) to (y, U_i, V_j), where V_j may be different from U_j. If they are different, their intersection should be empty.

In order to define the evaluation neighbourhood for each agent with respect to the state in question, we employ a technique inspired by the standard neighbourhood semantics [7]. We use a set of *neighbourhood functions*, determining the evaluation neighbourhood relative to both the given state and the corresponding agent. These functions need to be partial in order to render the semantics well-defined for the dynamic modalities in the system.

In Section 2 we define the syntax, structures, and semantics of our multi-agent logic of arbitrary public announcements, $APAL_{int}$, interpreted on topological spaces equipped with a set of neighbourhood functions. Without arbitrary announcements we get the logic PAL_{int}, and with neither arbitrary nor public announcements, the logic EL_{int}. In this section we also show some typical validities of the logic, and give a detailed example. In Section 3 we give axiomatizations for the logics: PAL_{int} extends EL_{int} and $APAL_{int}$ extends PAL_{int}. In Section 4 we demonstrate completeness for these logics. The completeness proof for the epistemic version of the logic, EL_{int}, is rather different from the completeness proof for the full logic $APAL_{int}$. We then compare our work to that of others (Section 5) and conclude.

2. THE LOGIC $APAL_{int}$

We define the syntax, structures, and semantics of our logic. From now on, *Prop* is a countable set of propositional variables and $\mathcal{A}$ a finite and non-empty set of agents.

2.1 Syntax

DEFINITION 1. *The language $\mathcal{L}_{APAL_{int}}$ is defined by*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid int(\varphi) \mid [\varphi]\varphi \mid \Box\varphi$$

where $p \in Prop$ and $i \in \mathcal{A}$. Abbreviations for the connectives $\vee$, $\rightarrow$ and $\leftrightarrow$ are standard, and $\bot$ is defined as abbreviation by $p \wedge \neg p$. We employ $\hat{K}_i$ for $\neg K_i\neg\varphi$, and $\Diamond\varphi$ for $\neg\Box\neg\varphi$. We denote the non-modal part of $\mathcal{L}_{APAL_{int}}$ (without the modalities K_i, int, $[\varphi]$ and $\Box$) by $\mathcal{L}_{Pl}$, the part without $\Box$ by $\mathcal{L}_{PAL_{int}}$, and the part without $\Box$ and $[\varphi]$ by $\mathcal{L}_{EL_{int}}$.

Necessity forms [10] allow us to select unique occurrences of a subformula in a given formula (unlike in uniform substitution). They will be used in the axiomatization (Section 3).

DEFINITION 2. *Let $\varphi \in \mathcal{L}_{APAL_{int}}$. The* necessity forms *are inductively defined as*

$$\xi(\sharp) := \sharp \mid \varphi \rightarrow \xi(\sharp) \mid K_i\xi(\sharp) \mid int(\xi(\sharp)) \mid [\varphi]\xi(\sharp).$$

It is not hard to see that each necessity form $\xi(\sharp)$ has a unique occurrence of $\sharp$. Given a necessity form $\xi(\sharp)$ and a formula $\varphi \in \mathcal{L}_{APAL_{int}}$, the formula obtained by replacing $\sharp$ by φ is denoted by $\xi(\varphi)$.

In the completeness proof (Section 4) we use a complexity measure on formulas based on the *size* and $\Box$-*depth* of formulas where the size of a formula is a weighted count of the number of symbols and $\Box$-*depth* counts the number of the $\Box$-modalities occurring in a formula. The measure was first introduced in [2].

DEFINITION 3. *The size $S(\varphi)$ of a formula $\varphi \in \mathcal{L}_{APAL_{int}}$ is defined as: $S(p) = 1$, $S(\neg\varphi) = S(\varphi) + 1$, $S(\varphi \wedge \psi) = S(\varphi) + S(\psi)$, $S(K_i\varphi) = S(\varphi) + 1$, $S(int(\varphi)) = S(\varphi) + 1$, $S([\varphi]\psi) = S(\varphi) + 4S(\psi)$, and $S(\Box\varphi) = S(\varphi) + 1$.*

The factor 4 in the clause for $[\varphi]\psi$ is to ensure Lemma 7. Although the choice of the number 4 might seem arbitrary, it is the smallest natural number guaranteeing the desired result (see the proof of Lemma 7).

DEFINITION 4. *The $\Box$-depth of a formula $\varphi \in \mathcal{L}_{APAL_{int}}$, denoted by $d(\varphi)$, is defined as: $d(p) = 0$, $d(\neg\varphi) = d(\varphi)$, $d(\varphi \wedge \psi) = max\{d(\varphi), d(\psi)\}$, $d(K_i\varphi) = d(\varphi)$, $d(int(\varphi)) = d(\varphi)$, $d([\varphi]\psi) = max\{d(\varphi), d(\psi)\}$, and $d(\Box\varphi) = d(\varphi) + 1$.*

We now define three order relations on $\mathcal{L}_{APAL_{int}}$ based on the size and $\Box$-depth of the formulas.

DEFINITION 5. *For any $\varphi, \psi \in \mathcal{L}_{APAL_{int}}$,*

- *$\varphi <^S \psi$ iff $S(\varphi) < S(\psi)$*
- *$\varphi <_d \psi$ iff $d(\varphi) < d(\psi)$*
- *$\varphi <^S_d \psi$ iff (either $d(\varphi) < d(\psi)$, or $d(\varphi) = d(\psi)$ and $S(\varphi) < S(\psi)$)*

We let $Sub(\varphi)$ denote the set of subformulas of a given formula φ.

LEMMA 6. *For any $\varphi, \psi \in \mathcal{L}_{APAL_{int}}$,*

1. *$<^S, <_d, <^S_d$ are well-founded strict partial orders between formulas in $\mathcal{L}_{APAL_{int}}$,*
2. *$\varphi \in Sub(\psi)$ implies $\varphi <^S_d \psi$,*
3. *$int(\varphi) <^S_d [\varphi]\psi$,*
4. *$\varphi \in \mathcal{L}_{PAL_{int}}$ iff $d(\varphi) = 0$,*
5. *$\varphi \in \mathcal{L}_{PAL_{int}}$ implies $[\varphi]\psi <^S_d \Box\psi$.*

LEMMA 7. *For any $\varphi, \psi, \chi \in \mathcal{L}_{APAL_{int}}$ and $i \in \mathcal{A}$,*

1. *$\neg[\varphi]\psi <^S_d [\varphi]\neg\psi$,*
2. *$int([\varphi]\psi) <^S_d [\varphi]int(\psi)$,*
3. *$K_i[\varphi]\psi <^S_d [\varphi]K_i\psi$,*
4. *$[\neg[\varphi]\neg int(\psi)]\chi <^S_d [\varphi][\psi]\chi$.*

PROOF. We only prove Lemma 7.4. The proof demonstrates why in the $[\varphi]\psi$ clause of Definition 3, 4 is the smallest natural number guaranteeing the result.

By Definition 3, we have that $S([\neg[\varphi]\neg int(\psi)]\chi) = S(\varphi) + 4S(\psi) + 4S(\chi) + 9$ and that $S([\varphi][\psi]\chi) = S(\varphi) + 4S(\psi) + 16S(\chi)$. As for any $\chi \in \mathcal{L}_{APAL_{int}}$, $1 \leq S(\chi)$, it follows that $4S(\chi)+9 \leq 4S(\chi)+9S(\chi) = 13S(\chi) < 16S(\chi)$. Further, we observe that $d([\neg[\varphi]\neg int(\psi)]\chi) = \max\{d(\varphi), d(\psi), d(\chi)\} = d([\varphi][\psi]\chi)$. (This is similar in the first three items.)

2.2 Background

In this section, we introduce the topological concepts that will be used throughout this paper. All the concepts in this section can be found in [9].

Definition 8. *A* topological space (X, τ) *is a pair consisting of a non-empty set X and a family τ of subsets of X satisfying $\emptyset \in \tau$ and $X \in \tau$, and closed under finite intersections and arbitrary unions.*

The set X is called the *space*. The subsets of X belonging to τ are called *open sets* (or *opens*) in the space; the family τ of open subsets of X is also called a *topology* on X. If for some $x \in X$ and an open $U \subseteq X$ we have $x \in U$, we say that U is an *open neighborhood* of x.

A point x is called an *interior point* of a set $A \subseteq X$ if there is an open neighborhood U of x such that $U \subseteq A$. The set of all interior points of A is called the *interior* of A and denoted by $Int(A)$. We can then easily observe that for any $A \subseteq X$, $Int(A)$ is the largest open subset of A.

Definition 9. *A family $B \subseteq \tau$ is called a* base *for a topological space (X, τ) if every non-empty open subset of X can be written as a union of elements of B.*

Given any family $\Sigma = \{A_\alpha \mid \alpha \in I\}$ of subsets of X, there exists a unique, smallest topology $\tau(\Sigma)$ with $\Sigma \subseteq \tau(\Sigma)$ [9, Th. 3.1]. The family $\tau(\Sigma)$ consists of $\emptyset$, X, all finite intersections of the A_α, and all arbitrary unions of these finite intersections. Σ is called a *subbase* for $\tau(\Sigma)$, and $\tau(\Sigma)$ is said to be generated by Σ. The set of finite intersections of members of Σ forms a base for $\tau(\Sigma)$.

2.3 Structures

In this section we define our multi-agent models based on topological spaces.

Definition 10. *Given a topological space (X, τ), a* neighbourhood function set Φ *on (X, τ) is a set of partial functions $\theta : X \rightharpoonup \mathcal{A} \to \tau$ such that for all $x, y \in Dom(\theta)$, for all $i \in \mathcal{A}$, and for all $U \in \tau$:*

1. $\theta(x)(i) \in \tau$,
2. $x \in \theta(x)(i)$,
3. $\theta(x)(i) \subseteq Dom(\theta)$,
4. *if* $y \in \theta(x)(i)$ *then* $\theta(x)(i) = \theta(y)(i)$,
5. $\theta|_U \in \Phi$,

where $\theta|_U$ is the partial function with $Dom(\theta|_U) = Dom(\theta) \cap U$ and $\theta|_U(x)(i) = \theta(x)(i) \cap U$. We call the elements of Φ neighbourhood functions.

Definition 11. *A* topological model with functions *(or in short, a* topo-model*) is a tuple $\mathcal{M} = (X, \tau, \Phi, V)$, where (X, τ) is a topological space, Φ a neighbourhood function set, and $V : Prop \to X$ a valuation function. We refer to the part $\mathcal{X} = (X, \tau, \Phi)$ without the valuation function as a* topo-frame.

A pair (x, θ) is a *neighbourhood situation* if $x \in Dom(\theta)$ and $\theta(x)(i)$ is called the *epistemic neighbourhood at x of agent i*. If (x, θ) is a neighbourhood situation in $\mathcal{M}$ we write $(x, \theta) \in \mathcal{M}$. Similarly, if (x, θ) is a neighbourhood situation in $\mathcal{X}$ we write $(x, \theta) \in \mathcal{X}$.

Lemma 12. *For any (X, τ, Φ) and $\theta \in \Phi$, $Dom(\theta) \in \tau$.*

2.4 Semantics

Definition 13. *Given a topo-model $\mathcal{M} = (X, \tau, \Phi, V)$ and a neighbourhood situation $(x, \theta) \in \mathcal{M}$, the semantics for the language $\mathcal{L}_{APAL_{int}}$ is defined recursively as:*

$$\begin{array}{lll}
\mathcal{M}, (x,\theta) \models p & \textit{iff} & x \in V(p) \\
\mathcal{M}, (x,\theta) \models \neg\varphi & \textit{iff} & \textit{not } \mathcal{M}, (x,\theta) \models \varphi \\
\mathcal{M}, (x,\theta) \models \varphi \wedge \psi & \textit{iff} & \mathcal{M}, (x,\theta) \models \varphi \textit{ and } \mathcal{M}, (x,\theta) \models \psi \\
\mathcal{M}, (x,\theta) \models K_i\varphi & \textit{iff} & (\forall y \in \theta(x)(i))(\mathcal{M}, (y,\theta) \models \varphi) \\
\mathcal{M}, (x,\theta) \models int(\varphi) & \textit{iff} & x \in Int[\![\varphi]\!]^\theta \\
\mathcal{M}, (x,\theta) \models [\varphi]\psi & \textit{iff} & \mathcal{M}, (x,\theta) \models int(\varphi) \Rightarrow \\
 & & \qquad \mathcal{M}, (x,\theta^\varphi) \models \psi \\
\mathcal{M}, (x,\theta) \models \Box\varphi & \textit{iff} & (\forall \psi \in \mathcal{L}_{PAL_{int}})(\mathcal{M}, (x,\theta) \models [\psi]\varphi)
\end{array}$$

where $p \in Prop$, $[\![\varphi]\!]^\theta = \{y \in Dom(\theta) \mid \mathcal{M}, (y,\theta) \models \varphi\}$ and $\theta^\varphi : X \rightharpoonup \mathcal{A} \to \tau$ such that $Dom(\theta^\varphi) = Int[\![\varphi]\!]^\theta$ and $\theta^\varphi(x)(i) = \theta(x)(i) \cap Int[\![\varphi]\!]^\theta$.

The *updated neighbourhood function* θ^φ is the restriction of θ to the open set $Int[\![\varphi]\!]^\theta$, i.e., for all $x \in X$, $\theta^\varphi(x)(i) = \theta|_{Int[\![\varphi]\!]^\theta}(x)(i)$.

A formula $\varphi \in \mathcal{L}_{APAL_{int}}$ is *valid in a topo-model* $\mathcal{M}$, denoted $\mathcal{M} \models \varphi$, iff $\mathcal{M}, (x,\theta) \models \varphi$ for all $(x,\theta) \in \mathcal{M}$; φ is *valid*, denoted $\models \varphi$, iff for all topo-models $\mathcal{M}$ we have $\mathcal{M} \models \varphi$. Soundness and completeness with respect to topo-models are defined as usual.

Let us now elaborate on the structure of topo-models and the above semantics we have proposed for $\mathcal{L}_{APAL_{int}}$. Given a topo-model (X, τ, Φ, V), the epistemic neighbourhoods of each agent at a given state x are determined by (partial) functions $\theta : X \rightharpoonup \mathcal{A} \to \tau$ assigning an open neighbourhood to the state in question for each agent. We allow for partial functions in Φ, and close Φ under taking restricted functions $\theta|_U$ where $U \in \tau$ (see Definition 10, condition 5), so that updated neighbourhood functions are guaranteed to be well-defined elements of Φ. As in the standard subset space semantics, by picking a neighbourhood situation (x, θ), we first localize our focus to an *open* subdomain, in fact to $Dom(\theta)$, including the state x and the epistemic neighbourhood of each agent at x determined by θ. Then the function $\theta(x)$ designates an epistemic neighbourhood for each agent i in $\mathcal{A}$. It is guaranteed that every agent i is assigned a neighbourhood by θ at every state x in $Dom(\theta)$, since each $\theta(x)$ is defined to be a *total* function from $\mathcal{A}$ to τ. Moreover, condition 2 of Definition 10 ensures that $\emptyset$ cannot be an epistemic neighbourhood, i.e., $\theta(x)(i) \neq \emptyset$ for all $x \in Dom(\theta)$. Finally, conditions 2 and 4 of Definition 10 make sure that the $S5$ axioms for each K_i are sound with respect to all topo-models.

We now provide some semantic results. As usual in the subset space setting, truth of non-modal formulas only depends on the state in question.

Proposition 14. *Give a topo-model $\mathcal{M} = (X, \tau, \Phi, V)$, neighbourhood situations $(x, \theta_1), (x, \theta_2) \in \mathcal{M}$, and a formula $\varphi \in \mathcal{L}_{Pl}$. Then $(x, \theta_1) \models \varphi$ iff $(x, \theta_2) \models \varphi$.*

Proposition 15. *Given $\mathcal{M} = (X, \tau, \Phi, V)$, $\theta \in \Phi$ and $\varphi \in \mathcal{L}_{APAL_{int}}$. Then $[\![int(\varphi)]\!]^\theta = Int[\![\varphi]\!]^\theta$.*

Proof.

$$\begin{array}{rcl}
[\![int(\varphi)]\!]^\theta & = & \{y \in Dom(\theta) \mid (y,\theta) \models int(\varphi)\} \\
 & = & \{y \in Dom(\theta) \mid y \in Int[\![\varphi]\!]^\theta\} \\
 & = & Int[\![\varphi]\!]^\theta \quad \text{(since } Int[\![\varphi]\!]^\theta \subseteq Dom(\theta)\text{)}
\end{array}$$

A corollary is that $Int[\![int(\varphi)]\!]^{\theta} = IntInt[\![\varphi]\!]^{\theta} = Int[\![\varphi]\!]^{\theta}$.

Proposition 16.

1. $\models [\varphi]\psi \leftrightarrow [int(\varphi)]\psi$

2. $\models (int(\varphi) \wedge \langle\varphi\rangle int(\psi)) \leftrightarrow \langle\varphi\rangle int(\psi)$

Proposition 17.

1. $[\![\psi]\!]^{\theta^{\varphi}} = [\![\langle\varphi\rangle\psi]\!]^{\theta}$

2. $\theta^{\varphi} = \theta^{int(\varphi)}$

3. $(\theta^{\varphi})^{\psi} = \theta^{\langle\varphi\rangle int(\psi)}$

2.5 Example

We illustrate our logic by a multi-agent version of Bjorndahl's convincing example in [6] about the jewel in the tomb. Indiana Jones (i) and Emile Belloq (e) are both scouring for a priceless jewel placed in a tomb. The tomb could either contain a jewel or not, the tomb could have been rediscovered in modern times or not, and (beyond [6]), the tomb could be in the Valley of Tombs in Egypt or not. The propositional variables corresponding to these propositions are, respectively, j, d, and t. We represent a valuation of these variables by a triple xyz, where $x, y, z \in \{0, 1\}$. Given carrier set $X = \{xyz \mid x, y, z \in \{0, 1\}\}$, the topology τ that we consider is generated by the base consisting of the subsets $\{000, 100, 001, 101\}$, $\{010\}$, $\{110\}$, $\{011\}$, $\{111\}$. The idea is that one can only conceivably know (or learn) about the jewel or the location, on condition that the tomb has been discovered. Therefore, $\{000, 100, 001, 101\}$ has no strict subsets besides empty set: if the tomb has not yet been discovered, no one can have any information about the jewel or the location.

A topo-model $\mathcal{M} = (X, \tau, \Phi, V)$ for this topology (X, τ) has Φ as the set of all neighbourhood functions that are partitions of X for both agents, and restrictions of these functions to open sets. A typical $\theta \in \Phi$ describes complete ignorance of both agents and is defined as $\theta(s)(i) = \theta(s)(e) = X$. This corresponds most to the situation described in [6]. A more interesting neighbourhood situation in this model is one wherein Indiana and Emile have different knowledge. Let us assume that Emile has the advantage over Indiana so far, as he knows the location of the tomb but Indiana doesn't. This is the θ' such that for all $x \in X$, $\theta'(x)(i) = X$ whereas the partition for Emile consists of sets $\{101, 100, 001, 000\}$, $\{111, 011\}$, $\{110, 010\}$, i.e., $\theta'(111)(e) = \{111, 011\}$, etc.

We now can evaluate what Emile knows about Indiana at 111, and confirm that this goes beyond Emil's initial epistemic neighbourhood. This situation however does not create any problems in our setting since Indiana's epistemic neighbourhoods will be determined relative to the states in Emile's initial neighbourhood. Firstly, Emile knows that the tomb is in the Valley of Tombs in Egypt

$$\mathcal{M}, (111, \theta') \models K_e t$$

and he also knows that Indiana does not know that

$$\mathcal{M}, (111, \theta') \models K_e \neg(K_i \neg t \vee K_i t)$$

The latter involves verifying $\mathcal{M}, (111, \theta') \models \hat{K}_i t$ and $\mathcal{M}, (111, \theta') \models \hat{K}_i \neg t$. And this is true because $\theta'(111)(i) = X$, and $000, 001 \in X$, and while $\mathcal{M}, (001, \theta') \models t$, we also have $\mathcal{M}, (000, \theta') \models \neg t$. We can also check that Emile knows that Indiana considers it possible that Emile doesn't know the tomb's location

$$\mathcal{M}, (111, \theta') \models K_e \hat{K}_i \neg(K_e t \vee K_e \neg t)$$

Announcements will change their knowledge in different ways. Consider the announcement of j. This results in Emile knowing everything but Indiana still being uncertain about the location.

$$\mathcal{M}, (111, \theta') \models [j](K_e(j \wedge d \wedge t) \wedge K_i(j \wedge d) \wedge \neg K_i(t \vee K_i \neg t))$$

Model checking this involves computing the epistemic neighbourhoods of both agents given by the updated neighbourhood function $(\theta')^j$ at 111. Observe that $Int[\![j]\!]^{\theta'} = \{111, 110\}$. Therefore, $(\theta')^j(111)(e) = Int[\![j]\!]^{\theta'} \cap \theta'(111)(e) = \{111\}$ and $(\theta')^j(111)(i) = Int[\![j]\!]^{\theta'} \cap \theta'(x)(i) = \{111, 110\}$.

There is an announcement after which Emile and Indiana know everything (for example the announcement of $j \wedge t$):

$$\mathcal{M}, (111, \theta) \models \Diamond(K_e(j \wedge d \wedge t) \wedge K_i(j \wedge d \wedge t))$$

As long as the tomb has not been discovered, nothing will make Emile (or Indiana) learn that it contains a jewel or where the tomb is located:

$$\mathcal{M} \models \neg d \rightarrow \Box(\neg(K_e j \vee K_e \neg j) \wedge \neg(K_e t \vee K_e \neg t))$$

3. AXIOMATIZATION

We now provide the axiomatizations of EL_{int}, PAL_{int}, and $APAL_{int}$, and prove their soundness and completeness with respect to the proposed semantics.

(P)	all instantiations of propositional tautologies	
(K-K)	$K_i(\varphi \rightarrow \psi) \rightarrow (K_i\varphi \rightarrow K_i\psi)$	
(K-T)	$K_i\varphi \rightarrow \varphi$	
(K-4)	$K_i\varphi \rightarrow K_iK_i\varphi$	
(K-5)	$\neg K_i\varphi \rightarrow K_i\neg K_i\neg\varphi$	
(int-K)	$int(\varphi \rightarrow \psi) \rightarrow (int(\varphi) \rightarrow int(\psi))$	
(int-T)	$int(\varphi) \rightarrow \varphi$	
(int-4)	$int(\varphi) \rightarrow int(int(\varphi))$	
(K_{int})	$K_i\varphi \rightarrow int(\varphi)$	
(R1)	$[\varphi]p \leftrightarrow (int(\varphi) \rightarrow p)$	
(R2)	$[\varphi]\neg\psi \leftrightarrow (int(\varphi) \rightarrow \neg[\varphi]\psi)$	
(R3)	$[\varphi](\psi \wedge \chi) \leftrightarrow [\varphi]\psi \wedge [\varphi]\chi$	
(R4)	$[\varphi]int(\psi) \leftrightarrow (int(\varphi) \rightarrow int([\varphi]\psi))$	
(R5)	$[\varphi]K_i\psi \leftrightarrow (int(\varphi) \rightarrow K_i[\varphi]\psi)$	
(R6)	$[\varphi][\psi]\chi \leftrightarrow [\neg[\varphi]\neg int(\psi)]\chi$	
(R7)	$\Box\varphi \rightarrow [\chi]\varphi$	where $\chi \in \mathcal{L}_{PAL_{int}}$
(DR1)	From φ and $\varphi \rightarrow \psi$, infer ψ	
(DR2)	From φ, infer $K_i\varphi$	
(DR3)	From φ, infer $int(\varphi)$	
(DR4)	From φ, infer $[\psi]\varphi$	
(DR5)	From $\xi([\psi]\chi)$ for all $\psi \in \mathcal{L}_{PAL_{int}}$, infer $\xi(\Box\chi)$	

Table 1: Axiomatizations EL_{int}, PAL_{int}, and $APAL_{int}$

Definition 18. *The axiomatization $APAL_{int}$ is given in Table 1. The axiomatization PAL_{int} is the one without (DR5) and (R7). We get EL_{int} if we further remove axioms (R1)-(R6) and the rule (DR4).*

The parts (DR1) to (DR5) are the *derivation rules* and the other parts are the *axioms*. A formula is a *theorem* of

$APAL_{int}$, notation $\vdash \varphi$, if it belongs to the smallest set of formulas containing the axioms and closed under the derivation rules. (Similarly for EL_{int} and PAL_{int}.)

Lemma 19. *Axiomatization $APAL_{int}$ satisfies substitution of equivalents. If $\vdash \varphi \leftrightarrow \psi$, then $\vdash \chi[p/\varphi] \leftrightarrow \chi[p/\psi]$.*

Proof. In the above, $\chi[p/\varphi]$ means uniform substitution of φ for p. The proof is not trivial but proceeds along similar lines as for public announcement logic, see [17].

Proposition 20. *$[\varphi]\bot \leftrightarrow \neg int(\varphi)$ is a theorem of $APAL_{int}$.*

Proposition 21. *$APAL_{int}$ is sound with respect to the class of all topo-models.*

Proof. Let $\mathcal{M} = (X, \tau, \Phi, V)$ be a topo-model, $(x, \theta) \in \mathcal{M}$ and $\varphi, \psi, \chi \in \mathcal{L}_{APAL_{int}}$. We show three cases.

($\mathbf{K}_{int}$) Suppose $(x, \theta) \models K_i\varphi$. This means, $(y, \theta) \models \varphi$ for all $y \in \theta(x)(i)$. Hence, $\theta(x)(i) \subseteq [\![\varphi]\!]^{\theta}$. By Definition 10, $\theta(x)(i)$ is an open neighbourhood of x, therefore we obtain $x \in Int[\![\varphi]\!]^{\theta}$, i.e., $(x, \theta) \models int(\varphi)$.

(R7) Let $\chi \in \mathcal{L}_{PAL_{int}}$ and suppose $(x, \theta) \models \Box\varphi$. By the semantics, we have $(x, \theta) \models \Box\varphi$ iff $(\forall\psi \in \mathcal{L}_{PAL_{int}})((x, \theta) \models [\psi]\varphi)$. Therefore, in particular, $(x, \theta) \models [\chi]\varphi$.

(DR5) Suppose $\xi([\psi]\chi)$ is valid for all $\psi \in \mathcal{L}_{PAL_{int}}$. The proof follows by induction on the complexity of $\xi(\sharp)$. In case $\xi(\sharp) = \sharp$, we have $\xi([\psi]\chi) = [\psi]\chi$. By assumption, we have that $[\psi]\chi$ is valid for all $\psi \in \mathcal{L}_{PAL_{int}}$. This implies $\mathcal{M}, (x, \theta) \models [\psi]\chi$ for all $\psi \in \mathcal{L}_{PAL_{int}}$, all topo-models $\mathcal{M}$, and $(x, \theta) \in \mathcal{M}$. Therefore, by the semantics, $\mathcal{M}, (x, \theta) \models \Box\chi$, i.e., $\mathcal{M}, (x, \theta) \models \xi(\Box\chi)$. All other, inductive, cases are elementary.

Corollary 22. *The axiomatizations EL_{int} and PAL_{int} are sound with respect to the class of all topo-models.*

4. COMPLETENESS

We now show completeness for EL_{int}, PAL_{int}, and $APAL_{int}$ with respect to the class of all topo-models. Completeness of EL_{int} is shown in a standard way via a canonical model construction and a Truth Lemma that is proved by induction on formula complexity. Completeness for PAL_{int} is shown by reducing each formula in $\mathcal{L}_{PAL_{int}}$ to an equivalent formula of $\mathcal{L}_{EL_{int}}$. The proof of the completeness for $APAL_{int}$ becomes more involved. Reduction axioms for public announcements no longer suffice in the $APAL_{int}$ case, and the inductive proof needs a subinduction where announcements are considered. Moreover, the proof system of $APAL_{int}$ has an infinitary derivation rule, namely the rule (DR5), and given the requirement of closure under this rule, the maximally consistent sets for that case are defined to be maximally consistent *theories* (see, Section 4.2). Lastly, the Truth Lemma requires the more complicated complexity measure on formulas defined in Section 2. There, we need to adapt the completeness proof of [2] to our setting.

4.1 Completeness of EL_{int} and PAL_{int}

For $\mathcal{L}_{EL_{int}}$ we define consistent and maximally consistent sets in the usual way, see e.g. [6] for details, and the multi-agent aspect does not complicate the definition. Let X^c be the set of all maximally consistent sets of EL_{int}. We define relations $\sim_i$ on X^c as $x \sim_i y$ iff $\forall\varphi \in \mathcal{L}_{EL_{int}}(K_i\varphi \in x$ iff $K_i\varphi \in y)$. Notice that the latter is equivalent to: $\forall\varphi \in \mathcal{L}_{EL_{int}}(K_i\varphi \in x$ implies $\varphi \in y)$ since K_i is an $S5$ modality. As each K_i is of $S5$ type, every $\sim_i$ is an equivalence relation, hence, it induces equivalence classes on X^c. Let $[x]_i$ denote the equivalence class of x induced by the relation $\sim_i$. Moreover, we define $\widehat{\varphi} = \{y \in X^c \mid \varphi \in y\}$. Observe that $x \in \widehat{\varphi}$ iff $\varphi \in x$.

Lemma 23 (Lindenbaum's Lemma). *Each consistent set can be extended to a maximally consistent set.*

Definition 24. *We define the canonical model $\mathcal{X}^c = (X^c, \tau^c, \Phi^c, V^c)$ as follows:*

- *X^c is the set of all maximally consistent sets;*
- *τ^c is the topological space generated by the subbase*

$$\Sigma = \{[x]_i \cap \widehat{int(\varphi)} \mid x \in X^c, \varphi \in \mathcal{L}_{EL_{int}} \text{ and } i \in \mathcal{A}\};$$

- *$x \in V^c(p)$ iff $p \in x$, for all $p \in Prop$;*
- *$\Phi^c = \{\theta^*|_U \mid U \in \tau^c\}$, where we define $\theta^* : X^c \to \mathcal{A} \to \tau^c$ as $\theta^*(x)(i) = [x]_i$, for $x \in X^c$ and $i \in \mathcal{A}$.*

Observe that, since $\widehat{int(\top)} = X^c$, we have $[x]_i \cap \widehat{int(\top)} = [x]_i \in \Sigma$ for each i. Therefore, each $[x]_i$ is an open subset of X^c. Moreover, the elements of Φ^c satisfy the required properties given in Definition 10.

Lemma 25 (Truth Lemma). *For every $\varphi \in \mathcal{L}_{EL_{int}}$ and for each $x \in X^c$, $\varphi \in x$ iff $\mathcal{X}^c, (x, \theta^*) \models \varphi$.*

Proof. Cases for the propositional variables and Booleans are straightforward. We only show the cases for K_i and int.

Case $\varphi := K_i\psi$

($\Rightarrow$) Suppose $K_i\psi \in x$ and let $y \in \theta^*(x)(i)$. Since $y \in \theta^*(x)(i) = [x]_i$, by definition of $\sim_i$, we have $K_i\psi \in y$. Then, by T-axiom for K_i, we obtain $\psi \in y$. Then, by IH, $\mathcal{X}^c, (y, \theta^*) \models \psi$. Therefore $\mathcal{X}^c, (x, \theta^*) \models K_i\psi$.

($\Leftarrow$) Suppose $K_i\psi \notin x$. Then, $\{K_i\gamma \mid K_i\gamma \in x\} \cup \{\neg\psi\}$ is a consistent set. We can then extend it to a maximally consistent set y. As $\{K_i\gamma \mid K_i\gamma \in x\} \subseteq y$, we have $y \in [x]_i$ meaning that $y \in \theta^*(x)(i)$. Moreover, since $\neg\psi \in y$, $\psi \notin y$. Therefore, we have a maximally consistent set $y \in \theta^*(x)(i)$ such that $\psi \notin y$. By (IH), $\mathcal{X}^c, (y, \theta^*) \not\models \psi$. Hence, $\mathcal{X}^c, (x, \theta^*) \not\models K_i\psi$.

Case $\varphi := int(\psi)$

($\Rightarrow$) Suppose $int(\psi) \in x$. Consider the set $[x]_i \cap \widehat{int(\psi)}$ for some $i \in \mathcal{A}$. Obviously, $x \in [x]_i \cap \widehat{int(\psi)}$ and $[x]_i \cap \widehat{int(\psi)}$ is open (since it is in Σ). Now let $y \in [x]_i \cap \widehat{int(\psi)}$. Since $y \in \widehat{int(\psi)}$, $int(\psi) \in y$. Then, by (int -T), since y is maximal consistent, we have $\psi \in y$. Thus, by IH, we have $(y, \theta^*) \models \psi$. Therefore, $y \in [\![\psi]\!]^{\theta^*}$. This implies $[x]_i \cap \widehat{int(\psi)} \subseteq [\![\psi]\!]^{\theta^*}$. And, since $x \in [x]_i \cap \widehat{int(\psi)} \in \tau^c$, we have $x \in Int[\![\psi]\!]^{\theta^*}$, i.e., $(x, \theta^*) \models int(\psi)$.

($\Leftarrow$) Suppose $(x, \theta^*) \models int(\psi)$, i.e., $x \in Int[\![\psi]\!]^{\theta^*}$. Recall that the set of finite intersections of the elements of Σ forms a base, which we denote by B_Σ, for τ^c. $x \in Int[\![\psi]\!]^{\theta^*}$ implies that there exists an open $U \in B_\Sigma$ such that $x \in U \subseteq [\![\psi]\!]^{\theta^*}$. Given the construction of B_Σ, U is of the form

$$U = \bigcap_{i \in I_1} [x_1]_i \cap \ldots \bigcap_{i \in I_n} [x_k]_i \cap \bigcap_{\eta \in \mathrm{Form}_{\mathrm{fin}}} \widehat{int(\eta)}$$

where $I_1, \ldots, I_n$ are finite subsets of $\mathcal{A}$, $x_1 \ldots x_k \in X^c$ and $\mathrm{Form}_{\mathrm{fin}}$ is a finite subset of $\mathcal{L}_{EL_{int}}$. Since int is a normal modality, we can simply write

$$U = \bigcap_{i\in I_1}[x_1]_i \cap \dots \bigcap_{i\in I_n}[x_k]_i \cap \widehat{int(\gamma)},$$

where $\bigwedge\limits_{\eta\in\mathrm{Form_{fin}}} \eta := \gamma$. Since x is in each $[x_j]_i$ with $1 \leq j \leq k$, we have $[x_j]_i = [x]_i$ for all such j. Therefore, we have

$$x \in U = (\bigcap_{i\in I}[x]_i) \cap \widehat{int(\gamma)} \subseteq [\![\psi]\!]^{\theta^*},$$

where $I = I_1 \cup \dots \cup I_n$.

This implies, for all $y \in (\bigcap\limits_{i\in I}[x]_i)$, if $y \in \widehat{int(\gamma)}$ then $\psi \in y$. From this, we can say $\bigcup\limits_{i\in I}\{K_i\sigma \mid K_i\sigma \in x\} \vdash int(\gamma) \to \psi$. Then, there is a finite subset $\Gamma \subseteq \bigcup\limits_{i\in I}\{K_i\sigma \mid K_i\sigma \in x\}$ such that $\vdash \bigwedge\limits_{\lambda\in\Gamma}\lambda \to (int(\gamma) \to \psi)$. It then follows:

1. $\vdash int(\bigwedge\limits_{\lambda\in\Gamma}\lambda \to (int(\gamma) \to \psi))$ (DR3)
2. $\vdash int(\bigwedge\limits_{\lambda\in\Gamma}\lambda) \to int(int(\gamma) \to \psi))$ (*int*-K) and (DR1)
3. $\vdash (\bigwedge\limits_{\lambda\in\Gamma} int(\lambda)) \to int(int(\gamma) \to \psi))$ (*int*-K)

Observe that each $\lambda \in \Gamma$ is of the form $K_j\alpha$ for some $K_j\alpha \in \bigcup\limits_{i\in I}\{K_i\sigma \mid K_i\sigma \in x\}$ and we have $\vdash K_i\varphi \leftrightarrow int(K_i\varphi)$. Therefore, $\vdash (\bigwedge\limits_{\lambda\in\Gamma}\lambda) \to int(int(\gamma) \to \psi))$. Thus, since $\bigwedge\limits_{\lambda\in\Gamma}\lambda \in x$ (by $\Gamma \subseteq x$), we have $int(int(\gamma) \to \psi)) \in x$. Then, by (*int*-K), (DR1) and since $\vdash int(int(\gamma)) \leftrightarrow int(\gamma)$ and $x \in \widehat{int(\gamma)}$ (i.e., $int(\gamma) \in x$) , we obtain $int(\psi) \in x$.

Theorem 26. *EL_{int} is complete with respect to the class of all topo-models.*

Theorem 27. *PAL_{int} is complete with respect to the class of all topo-models.*

Proof. This follows from Theorem 26 by reduction in a standard way. The occurrences of the modality *int* on the right-hand-side of the reduction axioms (axioms (R1)-(R6)) should not lead to any confusion: extending the complexity measure defined in [17, Definition 7.21 p. 187] to the language $\mathcal{L}_{PAL_{int}}$ by adding the same complexity measure for the modality *int* as for K_i gives us the desired result.

4.2 Completeness of $APAL_{int}$

We now reuse the technique of [2] in the setting of topological semantics. Given the closure requirement under derivation rule (DR5) it seems more proper to call maximally consistent sets of $APAL_{int}$ maximally consistent theories, as further explained below.

Definition 28. *A set x of formulas is called a* theory *iff $APAL_{int} \subseteq x$ and x is closed under (DR1) and (DR5). A theory x is said to be consistent iff $\perp \notin x$. A theory x is maximally consistent iff x is consistent and any set of formulas properly containing x is inconsistent.*

Observe that $APAL_{int}$ constitutes the smallest theory. Moreover, maximally consistent theories of $APAL_{int}$ posses the usual properties of maximally consistent sets:

Proposition 29. *For any maximally consistent theory x, $\varphi \notin x$ iff $\neg\varphi \in x$, and $\varphi \wedge \psi \in x$ iff $\varphi \in x$ and $\psi \in x$.*

In the setting of our axiomatization based on the infinitary rule (DR5), we will say that a set x of formulas is consistent iff there exists a consistent theory y such that $x \subseteq y$. Obviously, maximal consistent theories are maximal consistent sets of formulas. Under the given definition of consistency for sets of formulas, maximal consistent sets of formulas are also maximal consistent theories.

Definition 30. *Let $\varphi \in \mathcal{L}_{APAL_{int}}$ and $i \in \mathcal{A}$. Then $x + \varphi := \{\psi \mid \varphi \to \psi \in x\}$ and $K_i x := \{\varphi \mid K_i\varphi \in x\}$.*

Lemma 31. *For any theory x of $APAL_{int}$ and $\varphi \in \mathcal{L}_{APAL_{int}}$, $x + \varphi$ is a theory and it contains x and φ, and $K_i x$ is a theory.*

Lemma 32. *Let $\varphi \in \mathcal{L}_{APAL_{int}}$. For all theories x, $x + \varphi$ is consistent iff $\neg\varphi \notin x$.*

Proof. Let $\varphi \in \mathcal{L}_{APAL_{int}}$ and x be a theory. Then $\neg\varphi \in x$ iff $\varphi \to \perp \in x$ (as $\neg\varphi \leftrightarrow \varphi \to \perp$ is a theorem) iff $\perp \in x + \varphi$. Therefore, $x + \varphi$ is inconsistent iff $\neg\varphi \in x$, i.e., $x + \varphi$ is consistent iff $\neg\varphi \notin x$.

Lemma 33 (Lindenbaum's Lemma [1]). *Each consistent theory can be extended to a maximal consistent theory.*

Lemma 34. *If $K_i\varphi \notin x$, then there is a maximally consistent theory y such that $K_i x \subseteq y$ and $\varphi \notin y$.*

Proof. Let $\varphi \in \mathcal{L}_{APAL_{int}}$ and x be such that $K_i\varphi \notin x$. Thus, $\varphi \notin K_i x$. Hence, by Lemma 32, $K_i x + \neg\varphi$ is consistent. Then, by Lemma 33, there exists a maximally consistent set y such that $K_i x + \neg\varphi \subseteq y$. Therefore $K_i x \subseteq y$ and $\varphi \notin y$.

Lemma 35. *For all $\varphi \in \mathcal{L}_{APAL_{int}}$ and all maximally consistent theories x, $\Box\varphi \in x$ iff for all $\psi \in \mathcal{L}_{PAL_{int}}$, $[\psi]\varphi \in x$.*

Proof. Let $\varphi \in \mathcal{L}_{APAL_{int}}$ and x be a maximally consistent theory.

($\Rightarrow$) Suppose $\Box\varphi \in x$. Then, by (R7) and (DR1), we have $[\psi]\varphi \in x$ for all $\psi \in \mathcal{L}_{PAL_{int}}$.

($\Leftarrow$) Suppose $[\psi]\varphi \in x$ for all $\psi \in \mathcal{L}_{PAL_{int}}$. Consider the necessity form $\sharp$. By assumption, $\sharp([\psi]\varphi)$ for all $\psi \in \mathcal{L}_{PAL_{int}}$. Then, since x is closed under (DR5), $\sharp(\Box\varphi) \in x$, i.e., $\Box\varphi \in x$ as well.

The definition of *the canonical model for $APAL_{int}$* is the same as for EL_{int}, except that the maximally consistent sets are maximally consistent theories. We now come to the Truth Lemma for the logic $APAL_{int}$. Here we use the complexity measure $\psi <^S_d \varphi$.

Lemma 36 (Truth Lemma). *For every $\varphi \in \mathcal{L}_{APAL_{int}}$ and for each $x \in X^c$, $\varphi \in x$ iff $\mathcal{X}^c, (x, \theta^*) \models \varphi$.*

Proof. Let $\varphi \in \mathcal{L}_{APAL_{int}}$ and $x \in \mathcal{X}^c$. The proof is by $<^S_d$-induction on φ, where the case $\varphi = [\psi]\chi$ is proved by a subinduction on χ. We therefore consider 14 cases.

Case $\varphi := p$

$$\begin{array}{lll} x \in p & \text{iff} & x \in \nu^c(p) \\ & \text{iff} & (x, \theta^*) \models p \end{array}$$

Induction Hypothesis (IH): For all formulas $\psi \in \mathcal{L}_{APAL_{int}}$, if $\psi <^S_d \varphi$, then $\psi \in x$ iff $\mathcal{X}^c, (x, \theta^*) \models \psi$.

The cases negation, conjunction, and interior modality are as in Truth Lemma 25 for EL_{int}, where we observe that the subformula order is subsumed in the $<^S_d$ order (see Lemma 6.2). We proceed with the knowledge operator, i.e., case $\varphi := K_i\psi$, and then with the subinduction on χ for case announcement $\varphi := [\psi]\chi$, and finally with the case $\varphi := \Box\psi$.

Case $\varphi := K_i\psi$

This case is also similar to the one in Truth Lemma 25 for EL_{int}, however, using maximally consistent theories in the canonical model creates some differences. For the direction from left-to-right, see Truth Lemma 25. For ($\Leftarrow$), suppose $K_i\psi \notin x$. Then, by Lemma 34, there exists a maximally consistent theory y such that $K_ix \subseteq y$ and $\psi \notin y$. By $\psi <^S_d K_i\psi$ and (IH), $(y, \theta^*) \not\models \psi$. Since $K_ix \subseteq y$, we have $y \in [x]_i$ meaning that $y \in \theta^*(x)(i)$. Therefore, by the semantics, $\mathcal{X}^c, (x, \theta^*) \not\models K_i\psi$.

Case $\varphi := [\psi]p$

$$\begin{array}{llll}
[\psi]p \in x & \text{iff} & \mathit{int}(\psi) \to p \in x & \text{(R1)} \\
 & \text{iff} & \mathit{int}(\psi) \notin x \text{ or } p \in x & \text{Prop. 29} \\
 & \text{iff} & (x, \theta^*) \not\models \mathit{int}(\psi) \text{ or } (x, \theta^*) \models p & (*) \\
 & \text{iff} & (x, \theta^*) \models [\psi]p & \text{(R1)}
\end{array}$$

(*): By (IH), $\mathit{int}(\psi) <^S_d [\psi]p$ and $p <^S_d [\psi]p$ (Lemma 6.3 and Lemma 6.2).

Case $\varphi := [\psi]\neg\eta$ Use (R2) and (IH) and, by Lemma 6.3 and Lemma 7.1, $\mathit{int}(\psi) <^S_d [\psi]\neg\eta$ and $\neg[\psi]\eta <^S_d [\psi]\neg\eta$.

Case $\varphi := [\psi](\eta \wedge \sigma)$ Use (R3) and (IH), $[\psi]\eta <^S_d [\psi](\eta \wedge \sigma)$ and $[\psi]\sigma <^S_d [\psi](\eta \wedge \sigma)$.

Case $\varphi := [\psi]\mathit{int}(\eta)$ Use (R4) and (IH) and, by Lemmas 6.3, 7.2, $\mathit{int}(\psi) <^S_d [\psi]\mathit{int}(\eta)$ and $\mathit{int}([\psi]\eta) <^S_d [\psi]\mathit{int}(\eta)$.

Case $\varphi := [\psi]K_i\eta$ Use (R5) and (IH) and, by Lemmas 6.3, 7.3, $\mathit{int}(\psi) <^S_d [\psi]K_i\eta$ and $K_i[\psi]\eta <^S_d [\psi]K_i\eta$.

Case $\varphi := [\psi][\eta]\sigma$ Use (R6) and (IH) and, by Lemma 7.4, $[\neg[\psi]\neg \mathit{int}(\eta)]\sigma <^S_d [\psi][\eta]\sigma$.

Case $\varphi := [\psi]\Box\sigma$ For all $\eta \in \mathcal{L}_{PAL_{int}}$, $[\psi][\eta]\sigma <^S_d [\psi]\Box\sigma$, as $[\psi]\Box\sigma$ has one more $\Box$ than $[\psi][\eta]\sigma$. Therefore, it suffices to show $[\psi]\Box\sigma \in x$ iff $\forall\eta \in \mathcal{L}_{PAL_{int}}, [\psi][\eta]\sigma \in x$.

($\Leftarrow$) Consider the necessity form $[\psi]\sharp$ and assume that for all $\eta \in \mathcal{L}_{PAL_{int}}$, $[\psi][\eta]\sigma \in x$, i.e., for all $\eta \in \mathcal{L}_{PAL_{int}}$, $[\psi]\sharp([\eta]\sigma) \in x$. As x is closed under (DR5), we obtain $[\psi]\sharp(\Box\sigma) \in x$, i.e., $[\psi]\Box\sigma \in x$.

($\Rightarrow$) Suppose $[\psi]\Box\sigma \in x$. We have

$$\begin{array}{ll}
\vdash \Box\sigma \to [\eta]\sigma, \text{ for all } \eta \in \mathcal{L}_{PAL_{int}} & \text{(R7)} \\
\vdash [\psi](\Box\sigma \to [\eta]\sigma) \text{ for all } \eta \in \mathcal{L}_{PAL_{int}} & \text{(DR4)} \\
\vdash [\psi]\Box\sigma \to [\psi][\eta]\sigma, \text{ for all } \eta \in \mathcal{L}_{PAL_{int}} & \text{(DR1), (R1-R3)}
\end{array}$$

Therefore, for all $\eta \in \mathcal{L}_{PAL_{int}}, [\psi][\eta]\sigma \in x$. As $[\psi][\eta]\sigma <^S_d [\psi]\Box\sigma$ for all $\eta \in \mathcal{L}_{PAL_{int}}$, by (IH), we have for all $\eta \in \mathcal{L}_{PAL_{int}}, (x, \theta^*) \models [\psi][\eta]\sigma$. Then, by the semantics, we obtain (details omitted) that $(x, \theta^*) \models [\psi]\Box\sigma$.

Case $\varphi := \Box\psi$ Again note that for all $\eta \in \mathcal{L}_{PAL_{int}}$, $[\eta]\psi <^S_d \Box\psi$, as $\Box\psi$ has one more $\Box$ than $[\eta]\psi$ (see Lemma 6.4 and Lemma 6.5). Therefore, we obtain

$$\begin{array}{llll}
\Box\psi \in x & \text{iff} & (\forall\eta \in \mathcal{L}_{PAL_{int}})([\eta]\psi \in x) & \text{Lemma 35} \\
 & \text{iff} & (\forall\eta \in \mathcal{L}_{PAL_{int}})(x, \theta^*) \models [\eta]\psi & \text{(IH)} \\
 & \text{iff} & (x, \theta^*) \models \Box\psi & \text{semantics}
\end{array}$$

THEOREM 37. *$APAL_{int}$ is complete with respect to the class of all topo-models.*

PROOF. Let $\varphi \in \mathcal{L}_{APAL_{int}}$ such that $\not\vdash \varphi$, i.e., $\varphi \notin APAL_{int}$ (Recall that $APAL_{int}$ is the smallest theory). Then, by Lemma 32, $APAL_{int} + \neg\varphi$ is a consistent theory and, by Lemma 31, $\neg\varphi \in APAL_{int} + \neg\varphi$. By Lemma 33, the consistent theory $APAL_{int} + \neg\varphi$ can be extended to a maximally consistent theory y such that $APAL_{int} + \neg\varphi \subseteq y$. Since y is maximally consistent and $\neg\varphi \in y$, we obtain $\varphi \notin y$ (by Proposition 29). Then, by Lemma 36 (Truth Lemma), $\mathcal{X}^c, (y, \theta^*) \not\models \varphi$.

5. COMPARISON TO OTHER WORK

Multi-agent epistemic systems with subset space-like semantics have been proposed in [11, 12, 4, 18], however, none of these are concerned with arbitrary announcements. Our goal in this paper is not to provide a multi-agent generalization of SSL *per se*, but to work with the *effort-like* modality $\Box$ intended to capture the information change brought about by any announcements (subject to some restrictions) in a multi-agent setting and modelling it by way of "open-set shrinking" similar to the effort modality, rather than by deleting states or neighbourhoods, so that the intuitive link between the two becomes more transparent on a semantic level. In [3], Balbiani et al. proposed subset space semantics for arbitrary announcements, however, their approach does not go beyond the single-agent case and the semantics provided is in terms of model restriction. An unorthodox approach to multi-agent knowledge is proposed in [11, 12]. Roughly speaking, instead of having a knowledge modality K_i for each agent in his syntax, Heinemann uses additional operators to define K_i and his semantics only validate the $S4$-axioms for K_i. The necessitation rule for K_i does not preserve validity under the proposed semantics [11, 12]. In [18] a multi-agent semantics for knowledge is provided, but no announcements or further generalizations (unlike in their other, single-agent, work [19]), and not in a topological setting. Their use of partitions for each agent instead of a single neighbourhood is compatible with our requirement that all neighbourhoods for a given agent be disjoint. A further difference from the existing literature is that we restrict our attention to topological spaces and prove our results by means of topological tools.

We applied the new completeness proof for arbitrary public announcement logic of [2] to a topological setting. The canonical modal construction is as in [6] with some multi-agent modifications. The modality *int* in our system demands a different complexity measure in the Truth Lemma of the completeness proof than in [2].

6. CONCLUSIONS

We have proposed topological semantics for the multi-agent extensions of the public announcement logic of [6], and further extended the logic with arbitrary announcements. We showed topological completeness of these logics. Our work can be seen as a step toward discovering the interplay between dynamic epistemic logic and topological reasoning.

For further research, we envisage a **finitary** axiomatization for $APAL_{int}$ wherein the infinitary derivation rule (DR5) is replaced by a finitary rule. The obvious derivation rule would derive something after *any* announcement if it can be derived after announcing a fresh variable [1]. Under subset space semantics, it is unclear how to prove that this rule is sound.

We are still investigating expressivity and (un)decidability. If the logic $APAL_{int}$ is undecidable, this would contrast nicely with the undecidability of arbitrary public announcement logic. Otherwise, there may be interesting decidable versions when restricting the class of models to particular topologies.

The logic $APAL_{int}$ is also axiomatizable on the class where the K modalities have $S4$ properties, a result we have not reported in this paper for consistency of presentation. This class is of topological interest.

In our setup all agents have the same observational powers. If agents can have different observational powers, we can associate a topology with each agent and generalize the logic to an arbitrary *epistemic action* logic.

Furthermore, we would like to explore the exact difference between the effort modality and the arbitrary announcement modality (in the single agent case, see [16]) by constructing a topological model which distinguishes the two: a topological model might have more than epistemically definable opens with respect to the proposed semantics.

Acknowledgements

We thank Philippe Balbiani for various detailed suggestions over the past year on how to improve our single-agent and multi-agent results in subset space logic and topological logics. We have found him very supportive of our efforts. We also thank the TARK reviewers for their valuable comments. Hans van Ditmarsch is also affiliated to IMSc (Institute of Mathematical Sciences), Chennai, as research associate. We acknowledge support from European Research Council grant EPS 313360.

7. REFERENCES

[1] Balbiani, P., Baltag, A., van Ditmarsch, H., Herzig, A., Hoshi, T., and de Lima, T. (2008) ‘Knowable’ as ‘known after an announcement’. *The Review of Symbolic Logic*, **1**, 305–334.

[2] Balbiani, P. and van Ditmarsch, H. (2015) A simple proof of the completeness of APAL. *Studies in Logic*, **8 (1)**, 65–78.

[3] Balbiani, P., van Ditmarsch, H., and Kudinov, A. (2013) Subset space logic with arbitrary announcements. *Proc. of the 5th ICLA*, pp. 233–244, Springer.

[4] Baskent, C. (2007) *Topics in Subset Space Logic*. Master’s thesis, University of Amsterdam.

[5] Baskent, C. (2012) Public announcement logic in geometric frameworks. *Fundam. Inform.*, **118**, 207–223.

[6] Bjorndahl, A. (2013) Subset space public announcement logic revisited. *CoRR*, **abs/1302.4009**.

[7] Chellas, B. F. (1980) *Modal logic*. Cambridge University Press, Cambridge.

[8] Dabrowski, A., Moss, L. S., and Parikh, R. (1996) Topological reasoning and the logic of knowledge. *Ann. Pure Appl. Logic*, **78**, 73–110.

[9] Dugundji, J. (1965) *Topology*. Allyn and Bacon Series in Advanced Mathematics, Prentice Hall.

[10] Goldblatt, R. (1982) *Axiomatising the Logic of Computer Programming*. Springer-Verlag.

[11] Heinemann, B. (2008) Topology and knowledge of multiple agents. *Proc. of the 11th IBERAMIA*, pp. 1–10, Springer.

[12] Heinemann, B. (2010) Logics for multi-subset spaces. *Journal of Applied Non-Classical Logics*, **20**, 219–240.

[13] Moss, L. S. and Parikh, R. (1992) Topological reasoning and the logic of knowledge. *Proc. of the 4th TARK*, pp. 95–105, Morgan Kaufmann.

[14] Parikh, R., Moss, L., and Steinsvold, C. (2007) Topology and epistemic logic. *Handbook of Spatial Logics*, pp. 299–341.

[15] Plaza, J. A. (1989) Logics of public communications. *Proc. of the 4th ISMIS*, pp. 201–216, Oak Ridge National Laboratory.

[16] van Ditmarsch, H., Knight, S., and Özgün, A. (2014) Arbitrary announcements on topological subset spaces. *Proc. of the 12th EUMAS*, pp. 252–266, Springer.

[17] van Ditmarsch, H., van der Hoek, W., and Kooi, B. (2007) *Dynamic Epistemic Logic*, vol. 337 of *Synthese Library*. Springer.

[18] Wang, Y. N. and Ågotnes, T. (2013) Multi-agent subset space logic. *Proc. of the 23rd IJCAI*, pp. 1155–1161, IJCAI/AAAI.

[19] Wáng, Y. N. and Ågotnes, T. (2013) Subset space public announcement logic. *Proc. of 5th ICLA*, pp. 245–257, Springer.

Single-Peaked Consistency for Weak Orders Is Easy

Zack Fitzsimmons
College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY 14623, USA
zmf6921@rit.edu

ABSTRACT

In economics and social choice single-peakedness is one of the most important and commonly studied models for preferences. It is well known that single-peaked consistency for total orders is in P. However in practice a preference profile is not always comprised of total orders. Often voters have indifference between some of the candidates. In a weak preference order indifference must be transitive. We show that single-peaked consistency for weak orders is in P for three different variants of single-peakedness for weak orders. Specifically, we consider Black's original definition of single-peakedness for weak orders, Black's definition of single-plateaued preferences, and the existential model recently introduced by Lackner. We accomplish our results by transforming each of these single-peaked consistency problems to the problem of determining if a 0-1 matrix has the consecutive ones property.

Categories and Subject Descriptors

F.2.2 [**Theory of Computation**]: Analysis of Algorithms and Problem Complexity—*Nonnumerical Algorithms and Problems*; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent systems*

General Terms

Algorithms, Economics, Theory

Keywords

computational social choice, partial preferences, domain restrictions

1. INTRODUCTION

Single-peakedness is one of the most important and commonly examined domain restrictions on preferences in economics and social choice. The study of single-peaked preferences in computational social choice is often restricted to total orders, but in practical settings voters often have some degree of indifference in their preferences. This is seen in the online repository PrefLib, which contains several datasets comprised of voters with various degrees of partial preferences, many of which are weak orders [26]. Additionally, some election systems are defined for weak orders, e.g., the Kemeny rule [22] and the Schulze rule [30], or can be easily extended for weak orders.

Single-peaked preferences were introduced by Black [5] and they model the preferences of a collection of voters with respect to a one-dimensional axis, i.e., a total ordering of the candidates. Each voter in a single-peaked electorate has a single most preferred candidate (peak) on the axis and the farther that a candidate is from the voter's peak the less preferred they are by the voter. Black extended his model to single-plateaued preferences, which models the preferences of a collection of voters in a similar way, but allows voters to have multiple most preferred candidates (an indifference plateau) in their preferences [6, Chapter 5]. We mention that the definition of single-peaked preferences from Fishburn [18, Chapter 9] for weak orders is the same as Black's definition of single-plateaued preferences.

Elections where the voters have single-peaked preferences over the candidates have many desirable properties in economics and social choice, e.g., the majority relation is transitive [5] and there exist strategy-proof voting rules [28]. Additionally, computational problems often become easier when preferences are single-peaked. For example, when voters in an election have single-peaked (or even *nearly* single-peaked) preferences the complexity of determining if a manipulative action exists often becomes easy [17, 16] and determining the winner for Dodgson and Kemeny elections becomes easy [8] when it is Θ_2^p-complete in general [20, 21].

The problem of single-peaked consistency is to determine if an axis exists such that the preferences of a collection of voters are single-peaked. The first paper to computationally study single-peaked consistency for partial preferences was Lackner [24], where a partial order is said to be single-peaked with respect to an axis if it can be extended to a total order that is single-peaked with respect to that axis. For clarity we refer to this as existentially single-peaked, or ∃-single-peaked, throughout this paper. Lackner presents algorithms and complexity results for determining the ∃-single-peaked consistency for preference profiles of varying degrees of partial preferences, including top orders, weak orders, local weak orders, and partial orders. Lackner shows that if a given preference profile contains an implicitly specified total order (which is not guaranteed to exist) then ∃-single-peaked consistency for weak orders is in P [24]. Lackner also shows that the general case of ∃-single-peaked consistency for top orders (weak orders with all indifference between last-ranked candidates) is in P [24]. The complexity of the general case of ∃-single-peaked consistency for weak orders was explicitly left as the main open problem in Lackner [24] and we show in this paper that it is in P.

We show that an algorithm to determine if a 0-1 matrix has the consecutive ones property can be used to determine the single-peaked, single-plateaued, and ∃-single-peaked consistency for weak orders *without* requiring an implicitly specified total order. So given a preference profile of weak orders not only can we determine if it is single-peaked, single-plateaued, or ∃-single-peaked, we can find all consistent axes by using the PQ-tree algorithm for determining if a 0-1 matrix has the consecutive ones property [7]. This algorithm was previously used to determine the single-peaked consistency for total orders by Bartholdi and Trick [4] and to determine the single-crossing consistency for total orders by Bredereck et al. [9]. The model of single-crossing preferences is another domain restriction [27] and its corresponding consistency problem for total orders was first shown to be in P by Elkind et al. [12]. We also mention that after single-peaked consistency for total orders was shown to be in P, both Escoffier et al. [15] and Doignon and Falmagne [10] independently found faster direct algorithms.

This paper is organized as follows. In Section 2 we define the types of partial preferences studied, the different variants of single-peakedness, and the consecutive ones matrix problem. We present our results in Section 3, which is split into three sections with each corresponding to a different variant of single-peakedness. Section 3.1 contains our results for ∃-single-peaked preferences, Section 3.2 for single-plateaued preferences, and Section 3.3 for single-peaked preferences. In each of these sections we redefine the variant of single-peakedness using forbidden substructures and describe the transformation from its consistency problem to the problem of determining if a 0-1 matrix has the consecutive ones property. We conclude in Section 4 by summarizing our results and stating some possible directions for future work.

2. PRELIMINARIES

A *preference order*, v, is an ordering of the elements of a finite candidate set, C. A multiset of preference orders, V, is called a *preference profile.* (We sometimes refer to each v as a voter with a corresponding preference order.) A *partial order* is a transitive, reflexive, and antisymmetric binary relation on a set. A *weak order* is a partial order where the indifference relation is transitive, a *top order* is a weak order where all indifference is between candidates ranked last, and a *total order* is a partial order with no indifference between candidates.

Example 1. Given the set of candidates $\{a, b, c, d\}$, an example of a total order is $(a > b > d > c)$, an example of a weak order is $(a \sim c > d > b)$, and an example of a top order is $(a > c > b \sim d)$, where "$\sim$" is used to denote indifference between candidates.

We focus on weak orders since they model easily understood incompleteness in preferences. Voters are often not able to discern between two candidates or view them as truly equal. Allowing each voter to state a weak preference order still requires that they specify each candidate in their order, but gives them the ability to have multiple candidates at each position.

It is very natural for election systems to be defined for weak orders. The Kemeny rule and Schulze rule are defined for weak orders [22, 30] and clearly election systems based on pairwise comparisons (e.g., Copeland) can be used to evaluate a preference profile of partial votes. The Borda count can be extended for top orders [13] and a recent paper has even explored the complexity of the manipulation problem on such extensions to the Borda count and defined additional extensions for election systems to be defined on top orders [29].

2.1 Variants of Single-Peakedness

In our definitions of each variant of single-peakedness we refer to a total ordering of the set of candidates that each preference profile is consistent with as an axis A. Like Bartholdi and Trick [4], who were the first to show single-peaked consistency for total orders in P, we say that a preference order v is strictly increasing (decreasing) along a segment X of A if each candidate in X is strictly preferred to each candidate to its left (right) in X. Similarly, we say that a preference order is increasing (decreasing) along a segment X of A if each candidate in X is strictly preferred or ranked indifferent to each candidate to its left (right) in X. When we say that a preference order v is remaining constant along a segment, then all candidates in that segment are ranked indifferent to each other.

We begin our discussion of single-peaked preferences by stating the definition of single-peaked preferences for total orders. We use the definition found in the work by Bartholdi and Trick [4].

Definition 2. *A preference profile V of total orders is* single-peaked *with respect to an axis A if for every $v \in V$, A can be split at the most preferred candidate (peak) of v into two segments X and Y (one of which can be empty) such that v has strictly increasing preferences along X and v has strictly decreasing preferences along Y.*

We now define each of the three variants of single-peaked preferences for weak orders that we study in this paper and present an example of each in Figure 1.

2.1.1 Single-Peaked Preferences

Single-peaked preferences for weak orders can be defined in the same way as single-peaked preferences for total orders.

Definition 3. *A preference profile V of weak orders is* single-peaked *with respect to an axis A if for every $v \in V$, A can be split at the most preferred candidate (peak) of v into two segments X and Y (one of which can be empty) such that v has strictly increasing preferences along X and v has strictly decreasing preferences along Y.*

Notice that for a weak preference order to be single-peaked it must have a single most preferred candidate and can only contain indifference between at most two candidates at each position. Otherwise the segments X and Y referred to in Definition 3 would not be *strictly* increasing/decreasing. We define the corresponding problem of single-peaked consistency for weak orders below.

Given: A preference profile V of weak orders and a set of candidates C.

Question: Does there exist an axis A such that V is single-peaked with respect to A?

2.1.2 Single-Plateaued Preferences

A slightly weaker restriction than single-peakedness for weak orders is single-plateauedness [6, Chapter 5]. Single-peaked and single-plateaued preferences are closely related domain restrictions and Barberà [2] discusses how the amounts of indifference permitted in these restrictions impact their properties.

Building upon the definition for single-peaked preferences, we state a definition for single-plateaued preferences.

Definition 4. *A preference profile V of weak orders is* single-plateaued *with respect to an axis A if for every $v \in V$, A can be split into three segments X, Y, and Z (X and Z can each be empty) where v's most preferred candidates are Y, v has strictly increasing preferences along X, and v has strictly decreasing preferences along Z.*

We define the corresponding problem of single-plateaued consistency for weak orders below.

Given: A preference profile V of weak orders and a set of candidates C.

Question: Does there exist an axis A such that V is single-plateaued with respect to A?

2.1.3 Existentially Single-Peaked Preferences

So far we have considered the given preference orders as the true preferences of the voters. One approach to dealing with partial preferences is to assume that voters have an underlying total preference order and consider extensions of their preferences to total orders (see, e.g., [23]). This is the approach taken by Lackner for the existential model of single-peakedness [24].

Definition 5. *[24] A preference profile V of weak orders is* $\exists$-single-peaked *with respect to an axis A if for every $v \in V$, v can be extended to a total order v' such that the profile V' of total orders is single-peaked with respect to A.*

We can restate Definition 5 without referring to extensions to better see how it relates to single-peaked and single-plateaued preferences.

Observation 6. *A preference profile V of weak orders is* $\exists$-single-peaked *with respect to an axis A if and only if for every $v \in V$, A can be split into three segments X, Y, and Z (X and Z can each be empty) where v's most preferred candidates are Y, v has increasing preferences along X, and v has decreasing preferences along Z.*

We define the corresponding problem of $\exists$-single-peaked consistency for weak orders below.

Given: A preference profile V of weak orders and a set of candidates C.

Question: Does there exist an axis A such that V is $\exists$-single-peaked with respect to A?

Figure 1 illustrates an example of each variant of single-peakedness for weak orders described where each preference order is consistent with respect to the axis $A = a < d < b < e < c$. In Figure 1 the preference order $(b > d \sim e > c > a)$ is single-peaked, single-plateaued, and $\exists$-single-peaked. The preference order $(a \sim d > b > e > c)$ is single-plateaued and $\exists$-single-peaked, but not single-peaked since it has more than one most preferred candidate. The preference order $(c > b \sim e > d \sim a)$ is $\exists$-single-peaked and not single-plateaued or single-peaked since it is not strictly increasing to its most preferred candidate(s).

We conclude our discussion of these variants of single-peakedness for weak orders by stating several observations.

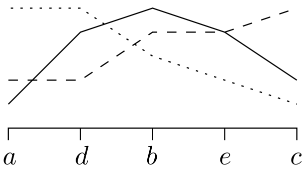


Figure 1: The solid line represents the single-peaked preference order $(b > d \sim e > c > a)$, the dotted line represents the single-plateaued preference order $(a \sim d > b > e > c)$, and the dashed line represents the $\exists$-single-peaked preference order $(c > b \sim e > d \sim a)$.

First we show that there exists an $\exists$-single-peaked consistent preference profile that does not have a transitive majority relation. We say that a majority relation is transitive if when $x > y$ and $y > z$ by majority, then $x > z$ by majority. Note that single-peaked and single-plateaued preferences both have transitive majority relations [5, 6].

Consider the preference profile V comprised of the following five voters from Table 9.1 in Fishburn [18].

$$\begin{array}{ll} v_1 & (b > a > c) \\ v_2, v_3 & (c > b > a) \\ v_4, v_5 & (a > b \sim c) \end{array}$$

When we evaluate this preference profile under the simple majority rule where $x > y$ by simple majority if more voters state $x > y$ than $y > x$, then V has the majority cycle $a > c > b > a$ [18]. Clearly V is $\exists$-single-peaked consistent with respect to the axis $A = a < b < c$, so we can make the following observation.

Observation 7. *There exists a preference profile of weak orders that is $\exists$-single-peaked and does not have a transitive majority relation.*

The existential model for single-peakedness considers the existence of a single extension of the preferences of all of the voters to total orders. We briefly consider the case where all extensions to total orders must be single-peaked and make two observations.

Observation 8. *If a preference profile of weak orders is single-peaked then all extensions of the preferences to total orders are also single-peaked.*

Observation 9. *If a preference profile of weak orders is single-plateaued and each preference order has at most two most preferred candidates, then all extensions of the preferences to total orders are single-peaked.*

2.2 Consecutive-Ones Matrices

All of our polynomial-time results are due to transformations to the following problem of determining if a 0-1 matrix has the consecutive ones property.

Given: A 0-1 matrix M.

Question: Does there exist a permutation of the columns of M such that in each row all of the 1's are consecutive?

The above problem was shown to be in P by Fulkerson and Gross [19]. Booth and Lueker [7] improved on this result by finding a linear-time algorithm through the development and use of the novel PQ-tree data structure, which contains all possible permutations of the columns of a matrix such that all of the 1's are consecutive in each row.

3. RESULTS

The following three sections consist of our results and they are structured as follows. We examine each variant of single-peakedness starting with the weakest restriction and ending with the strongest. When we examine each restriction we present an alternative definition of the variant of single-peakedness using forbidden substructures and the transformation to the problem of determining if a 0-1 matrix has the consecutive ones property.

3.1 Existentially Single-Peaked Consistency

The most general of the three variants mentioned in Section 2.1 is the model of $\exists$-single-peaked preferences. The construction and corresponding proof will be the basis for showing that single-peaked and single-plateaued consistency for weak orders are each also in P.

Given an axis A and a preference order v, if v is $\exists$-single-peaked with respect to A then v cannot have strictly decreasing and then strictly increasing preferences with respect to A. Following the terminology used by Lackner [24], we refer to this as a v-valley.

Definition 10. *A preference order v over a candidate set C contains a* v-valley *with respect to an axis A if there exist candidates $a, b, c \in C$ such that $a < b < c$ in A and $(a > b)$ and $(c > b)$ in v.*

Using the v-valley substructure we can state the following lemma, which will simplify our argument used in the proof of Theorem 14.

Lemma 11. *[24] Let V be a preference profile of weak orders. V is $\exists$-single-peaked with respect to an axis A if and only if no preference order $v \in V$ contains a v-valley with respect to A.*

To construct a matrix from a preference profile of weak orders, we apply essentially the same transformation as used Bartholdi and Trick [4] for total orders (see Example 13). We describe the construction below.

Construction 12. *Let V be a preference profile of weak orders over candidate set C. For each $v \in V$ construct a $(\|C\| - 1) \times \|C\|$ matrix X_v. Each column of X_v corresponds to a candidate in C. For each candidate $c \in C$ let k be the number of candidates that are strictly preferred to c in v and let the corresponding column in matrix X_v contain k 0's starting at row one, with the remaining entries filled with 1's. All $\|V\|$ of the matrices are row-wise concatenated to yield the $(\|V\| \cdot (\|C\| - 1)) \times \|C\|$ matrix X.*

The main difference in our construction is that we have one fewer row in each of the individual preference matrices. In the construction used by Bartholdi and Trick [4], given a preference order v over a set of candidates C, for all $a, b \in C$, $(a > b)$ in v if and only if the number of 1's in the column corresponding to a is greater than the number of 1's in the column corresponding to b in v's corresponding individual preference matrix. Notice that this still holds for our construction.

The polynomial-time results for $\exists$-single-peaked consistency for weak orders and local weak orders proved in Lackner [24] require that the given preference profile contains a *guiding order*, i.e., an implicitly specified total order. Given a preference profile V, a guiding order can be constructed iteratively in the following way. If there exists a $v \in V$ such that the last ranked candidate in v is not ranked indifferently with any other candidate, then that candidate is appended to the top of the guiding order. This is then repeated on the preference profile restricted to the candidates not yet added to the guiding order until either the guiding order is a total order or there is no $v \in V$ with a unique last ranked candidate, the case where no guiding order exists [24]. Observe that if a given preference profile is $\exists$-single-peaked then it remains $\exists$-single-peaked if a guiding order is added as an additional preference order [24]. It is important to point out that our results do not depend on the existence of a guiding order in a preference profile. Below we show how Construction 12 is applied to a preference profile of weak orders that is $\exists$-single-peaked.

Example 13. Consider the preference profile V that consists of the preference orders v and w. Let the preference order v be $(a \sim c > b > e \sim d > f)$ and the preference order w be $(a > b > c > e \sim d > f)$. Notice that V does not contain a guiding order, which is required by the polynomial-time algorithm for weak orders found in Lackner [24].

$$X_v = \begin{array}{cccccc} a & b & c & d & e & f \\ \hline 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{array} \qquad X_w = \begin{array}{cccccc} a & b & c & d & e & f \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{array}$$

We then row-wise concatenate X_v and X_w to construct X.

$$X = \begin{array}{cccccc} a & b & c & d & e & f \\ \hline 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{array} \qquad X' = \begin{array}{cccccc} b & a & c & d & e & f \\ \hline 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{array}$$

Next, we permute the columns of X so that in each row all of the 1's are consecutive to yield X'. Observe that V is $\exists$-single-peaked with respect to $b < a < c < d < e < f$, the ordering of the columns of X' as its axis.

We now show that $\exists$-single-peaked consistency for weak orders and the problem of determining if the constructed 0-1 matrix has the consecutive ones property are equivalent using Lemma 11 and Construction 12.

Theorem 14. *A preference profile V of weak orders is $\exists$-single-peaked consistent if and only if the matrix X, constructed using Construction 12, has the consecutive ones property.*

Proof. Let V be a preference profile of weak orders. Essentially the same argument as used by Bartholdi and Trick [4] holds.

If V is $\exists$-single-peaked with respect to an axis A then by Lemma 11 we know that no preference order $v \in V$ contains a v-valley with respect to A. When the columns of the matrix X are permuted to correspond to the axis A no row will contain the sequence $\cdots 1 \cdots 0 \cdots 1 \cdots$ since this corresponds to a preference order that strictly decreases and then strictly increases along the axis A (a v-valley). Therefore X has the consecutive ones property.

For the other direction suppose that V is not $\exists$-single-peaked, then by Lemma 11 we know that for every possible axis there exists a preference order $v \in V$ such that v contains a v-valley with respect to that axis. So every permutation of the columns of X will correspond to an axis where some preference order has a v-valley. As stated in the other direction, a v-valley corresponds to a row containing the sequence $\cdots 1 \cdots 0 \cdots 1 \cdots$ so clearly X does not have the consecutive ones property.

The only difference from the argument used by Bartholdi and Trick [4] for total orders is that in our case the preference orders can remain constant at the peak and at points on either side of the peak. The same argument still applies since by Lemma 11 the absence of v-valleys with respect to an axis is equivalent to a profile of weak orders being $\exists$-single-peaked with respect to that axis. $\Box$

Corollary 15. *$\exists$-Single-peaked consistency for weak orders is in* P.

3.2 Single-Plateaued Consistency

Single-plateaued preferences are a much more restrictive model than $\exists$-single-peaked preferences since they are essentially single-peaked except that each preference order can have multiple most preferred candidates [6, Chapter 5].

Since a preference order must be strictly increasing and then strictly decreasing with respect to an axis (excluding its most preferred candidates) we can again use the v-valley substructure. However we will need another substructure to prevent two candidates that are ranked indifferent in a voter's preference order from appearing on the same side of that voter's peak (plateau).

Definition 16. *A preference order v over a candidate set C contains a* nonpeak plateau *with respect to A if there exist candidates $a, b, c, \in C$ such that $a < b < c$ in A and either $(a > b \sim c)$ or $(c > b \sim a)$ in v.*

We use the v-valley and nonpeak plateau substructures to state the following lemma.

Lemma 17. *Let V be a preference profile of weak orders. V is single-plateaued with respect to an axis A if and only if no preference order $v \in V$ contains a v-valley with respect to A and no preference order $v \in V$ contains a nonpeak plateau with respect to A.*

Proof. Let C be a candidate set, V be a preference profile of weak orders, and A be an axis.

If V is single-plateaued with respect to A then for every preference order $v \in V$, A can be split into segments X, Y, and Z such that v is strictly increasing along X, remaining constant along Y, and strictly decreasing along Z. Since v is only ever strictly decreasing along Z and Z is the rightmost segment of A, v cannot contain a v-valley with respect to A. For a nonpeak plateau to exist with respect to A there must exist candidates $a, b, c \in C$ such that $a < b < c$ in A and either $(a > b \sim c)$ or $(c > b \sim a)$ in v.

We first consider the case of $(a > b \sim c)$ in v. Since a is strictly preferred to b and c in v and both b and c are to the right of a on the axis we know that both b and c must be in segment Z. However, v is strictly decreasing along Z, so v cannot have a nonpeak plateau of this form.

We now consider the case of $(c > b \sim a)$ in v. Since c is strictly preferred to a and b in v and both a and b are to the left of c on the axis we know that both a and b must be in segment X. However, v is strictly increasing along X, so v cannot have a nonpeak plateau of this form.

For the other direction we consider the case when no preference order $v \in V$ contains a v-valley with respect to A and no preference order $v \in V$ contains a nonpeak plateau with respect to A.

Since no preference order $v \in V$ contains a v-valley with respect to A, we know from Lemma 11 that V is $\exists$-single-peaked with respect to A. Since we also know that no preference order $v \in V$ contains a nonpeak plateau with respect to A it is easy to see that V is single-plateaued with respect to A. $\Box$

Since the nonpeak plateau substructure is needed in addition to the v-valley substructure, we need to extend Construction 12 so that if a preference order contains a nonpeak plateau with respect to an axis A, then when the columns of its corresponding preference matrix are permuted according to A the matrix will contain a row with the sequence $\cdots 1 \cdots 0 \cdots 1 \cdots$.

Notice that if a preference order ranks three candidates indifferent to each other below its peak (plateau) that it will have a nonpeak plateau with respect to *every* possible axis. To handle this case in the extension to Construction 12 we need to ensure that its corresponding preference matrix will contain a row with the sequence $\cdots 1 \cdots 0 \cdots 1 \cdots$ for every permutation of its columns.

Construction 18. *Let V be a preference profile of weak orders over candidate set C. For each $v \in V$ construct a $(\|C\| - 1) \times \|C\|$ matrix X_v. Each column of X_v corresponds to a candidate in C. For each candidate $c \in C$ let k be the number of candidates that are strictly preferred to c in v and let the corresponding column in matrix X_v contain k 0's starting at row one, with the remaining entries filled with 1's (as in Construction 12). The following extensions to Construction 12 ensure that if v has nonpeak plateau with respect to an axis A then when the columns of X_v are permuted according to A it will not have consecutive ones in rows.*

For each pair of candidates $a, b \in C$ such that $(a \sim b)$ in v, they are not the most preferred candidates in v, and there is no candidate $c \in C - \{a, b\}$ such that v is indifferent among a, b, and c, then append three additional rows to the matrix X_v where the column corresponding to a is $\begin{bmatrix}0 & 1 & 1\end{bmatrix}'$, the column corresponding to b is $\begin{bmatrix}1 & 1 & 0\end{bmatrix}'$, each column corresponding to a candidate strictly preferred to a and b is $\begin{bmatrix}1 & 1 & 1\end{bmatrix}'$, and each column corresponding to a remaining candidate is $\begin{bmatrix}0 & 0 & 0\end{bmatrix}'$.

If there exist three candidates $a, b, c \in C$ such that $(a \sim b \sim c)$ in v and they are not the most preferred candidates in v, then output a matrix that has no solution.

After constructing an X_v matrix for each $v \in V$, all $\|V\|$ of the matrices are row-wise concatenated to yield a matrix X, except in the case where the input resulted in a matrix with no solution.

We now show how Construction 18 is applied to a preference profile of weak orders that is single-plateaued.

Example 19. We consider the same preference profile as in Example 13 and we bold the additional rows in this example. Let the preference order v be $(a \sim c > b > e \sim d > f)$ and the preference order w be $(a > b > c > e \sim d > f)$.

$$X_v = \begin{array}{c} \begin{matrix} a & b & c & d & e & f \end{matrix} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \end{bmatrix} \end{array} \quad X_w = \begin{array}{c} \begin{matrix} a & b & c & d & e & f \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \end{bmatrix} \end{array}$$

We then row-wise concatenate X_v and X_w to construct X.

$$X = \begin{array}{c} \begin{matrix} a & b & c & d & e & f \end{matrix} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \end{bmatrix} \end{array} \quad X' = \begin{array}{c} \begin{matrix} e & b & a & c & d & f \end{matrix} \\ \begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} \end{bmatrix} \end{array}$$

Next, we permute the columns of X such that in each row all of the ones are consecutive to yield X'. Observe that V is single-plateaued with respect to this new ordering $e < b < a < c < d < f$ as its axis. Also notice that an axis containing d and e adjacent to each other (as seen in Example 13) would not correspond to an ordering of the columns of X with consecutive ones in rows due to the additional rows from the extensions made to Construction 12 in Construction 18.

Construction 12 ensures that no preference order contains a v-valley and the extensions made in Construction 18 ensure that no preference order contains a nonpeak plateau. So the proof of the following theorem uses a similar argument to the proof of Theorem 14. Now the presence of v-valleys *or* nonpeak plateaus, not just v-valleys, is equivalent to a row containing the sequence $\cdots 1 \cdots 0 \cdots 1 \cdots$.

Theorem 20. *A preference profile V of weak orders is single-plateaued consistent if and only if the matrix X, constructed using Construction 18, has the consecutive ones property.*

Proof. Let V be a preference profile of weak orders. We extend the argument used by Bartholdi and Trick [4] and the proof of Theorem 14 except in this case we use Lemma 17 instead of Lemma 11.

If V is single-plateaued with respect to an axis A then by Lemma 17 we know that no $v \in V$ contains a v-valley with respect to A and no $v \in V$ contains a nonpeak plateau with respect to A. When the columns of the matrix X are permuted to correspond to the axis A no row will contain the sequence $\cdots 1 \cdots 0 \cdots 1 \cdots$ since this would correspond to a preference order that strictly decreases and then strictly increases along the axis A (a v-valley) or it would correspond to a preference order that has two candidates ranked indifferent appearing on the same side of its peak (a nonpeak plateau). Therefore X has the consecutive ones property.

If V is not single-plateaued then we know from Lemma 17 that for every possible axis there exists a preference order $v \in V$ such that v contains a v-valley or v contains a nonpeak plateau with respect to that axis. So every permutation of the columns of X will correspond to an axis where a preference order has a v-valley or a nonpeak plateau. As stated in the other direction, the presence of a v-valley or a nonpeak plateau corresponds to a row containing the sequence $\cdots 1 \cdots 0 \cdots 1 \cdots$. Therefore X does not have the consecutive ones property. □

Corollary 21. *Single-plateaued consistency for weak orders is in* P.

3.3 Single-Peaked Consistency

We now present our results for the strongest domain restriction on weak orders that we examine, single-peaked preferences. Recall that a preference order is single-peaked with respect to an axis A if it is strictly increasing to a single most preferred candidate (peak) and then strictly decreasing with respect to A. So we again use the v-valley substructure, but like the previous case of single-plateaued preferences we need an additional substructure. Even if no preference order has a v-valley with respect to A it may not be single-peaked because it is indifferent between two candidates on the same side of its peak or has more than one most preferred candidate.

We can handle the first condition just mentioned with the nonpeak plateau substructure used in Section 3.2, but the second condition requires us to view *any* plateau as a forbidden substructure.

Definition 22. *A preference order v over a candidate set C contains a* plateau *with respect to an axis A if there exist candidates $a, b \in C$ such that a and b are adjacent in A and $(a \sim b)$ in v.*

We can now use the plateau substructure and the v-valley substructure to state the following lemma.

Lemma 23. *Let V be a preference profile of weak orders. V is single-peaked with respect to an axis A if and only if no preference order $v \in V$ contains a v-valley with respect to A and no preference order $v \in V$ contains a plateau with respect to A.*

Proof. Let C be a candidate set, V be a preference profile of weak orders, and A be an axis.

If V is single-peaked with respect to A then clearly V is also single-plateaued with respect to A. So by Lemma 17 we know that no preference order $v \in V$ contains a v-valley with respect to A and no preference order $v \in V$ contains a nonpeak plateau with respect to A. Since V is single-peaked we also know that no preference order $v \in V$ has more than one most preferred candidate so clearly no preference order $v \in V$ contains a plateau with respect to A.

For the other direction we consider the case when no preference order $v \in V$ contains a v-valley with respect to A and no preference order $v \in V$ contains a plateau with respect to A.

Since no preference order $v \in V$ contains a v-valley with respect to A we know from Lemma 11 that V is $\exists$-single-peaked with respect to A. Since we also know that no preference order $v \in V$ contains a plateau with respect to A it is easy to see that V is single-peaked with respect to A. □

We extend Construction 18 so that if a preference order contains a plateau with respect to an axis A, then when the columns of its preference matrix are permuted according to A the matrix will contain the sequence $\cdots 1 \cdots 0 \cdots 1 \cdots$. Since Construction 18 already ensures this for the case of nonpeak plateaus, our extended construction below only needs to add a condition for plateaus that contain the most preferred candidates in a given preference order.

Construction 24. *Follow Construction 18 except add the following condition while constructing a preference matrix X_v for each preference order $v \in V$.*

If there exist two candidates $a, b \in C$ such that $(a \sim b)$ in v and they are the most preferred candidates in v, then output a matrix that has no solution.

Clearly the extension to Construction 18 above ensures that if there are multiple most preferred candidates in a preference order then the preference matrix constructed from that order does not have the consecutive ones property.

When a preference order has a unique most preferred candidate and is single-plateaued, it is clearly also single-peaked. Construction 24 ensures that no preference order contains more than one most preferred candidate the same way that Construction 18 ensures that no preference order contains three or more candidates that are all ranked indifferent to each other and that are not the most preferred candidates, since this always results in a nonpeak plateau. So the proof of the following theorem follows from the proof of Theorem 20, but using Lemma 23 instead of Lemma 17.

Theorem 25. *A preference profile V of weak orders is single-peaked consistent if and only if the matrix X, constructed using Construction 24 has the consecutive ones property.*

Corollary 26. *Single-peaked consistency for weak orders is in* P.

4. CONCLUSIONS AND FUTURE WORK

We presented three different variants of single-peaked preferences for weak orders and showed that each of their corresponding consistency problems are in P. Since we accomplished this by using transformations to the problem of determining if a 0-1 matrix has the consecutive ones property we are able to apply the PQ-tree algorithm from Booth and Lueker [7]. Using this algorithm we can actually go further than just determining the consistency problem for each of these variants and find *all* consistent axes. An interesting open direction is how the consecutive ones matrix problem relates to other domain restrictions and what benefits there are to having all consistent axes for a given preference profile.

The existential approach introduced by Lackner for single-peaked preferences [24] has been recently applied to other domain restrictions. The model of single-crossing preferences [27] was studied in the existential model by Elkind et al. [11] and the model of top-monotonic preferences [3] was studied in the existential model by Aziz [1]. An interesting direction for future work would be to apply the existential model to other domain restrictions.

Single-peaked preferences are studied because they are a simply stated and important domain restriction that gives insight into how the voters view the candidates and elections with single-peaked voters have nice properties. However, experimental study suggests that in real-world settings voters are often not single-peaked [25], but in this study the single-peaked results only used Black's definition for total orders. It would be interesting to see if real-world datasets of weak orders contain voters that are single-peaked, single-plateaued, or $\exists$-single-peaked.

In single-peaked and nearly single-peaked elections computational problems often become easier [17, 16]. As mentioned by Lackner [24] an important open problem is to determine what computational benefits are gained when a preference profile is $\exists$-single-peaked or even nearly $\exists$-single-peaked. There are several different types of nearly single-peakedness and determining if a given preference profile is nearly single-peaked with respect to a certain distance measure is an interesting computational problem [14]. It would be interesting to see how preference profiles of weak orders impact the complexity of nearly single-peakedness or, as also mentioned by Lackner [24], nearly single-peakedness in the existential model.

5. ACKNOWLEDGMENTS

The author thanks Edith Hemaspaandra, Martin Lackner, David Narváez, Stanisław Radziszowski, and the anonymous referees for their helpful suggestions and feedback. This work was supported in part by NSF grant no. CCF-1101452 and a National Science Foundation Graduate Research Fellowship under NSF grant no. DGE-1102937.

6. REFERENCES

[1] H. Aziz. Testing top monotonicity. Technical Report arXiv:1403.7625 [cs.GT], arXiv.org, June 2014.

[2] S. Barberà. Indifferences and domain restrictions. *Analyse & Kritik*, 29(2):146–162, 2007.

[3] S. Barberà and B. Moreno. Top monotonicity: A common root for single peakedness, single crossing and the median voter result. *Games and Economic Behavior*, 73(2):345–359, 2011.

[4] J. Bartholdi, III and M. Trick. Stable matching with preferences derived from a psychological model. *Operations Research Letters*, 5(4):165–169, 1986.

[5] D. Black. On the rationale of group decision-making. *Journal of Political Economy*, 56(1):23–34, 1948.

[6] D. Black. *The Theory of Committees and Elections.* Cambridge University Press, 1958.

[7] K. Booth and G. Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *Journal of Computer and System Sciences*, 13(3):335–379, 1976.

[8] F. Brandt, M. Brill, E. Hemaspaandra, and L. A. Hemaspaandra. Bypassing combinatorial protections: Polynomial-time algorithms for single-peaked electorates. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pages 715–722, July 2010.

[9] R. Bredereck, J. Chen, and G. Woeginger. A characterization of the single-crossing domain. *Social Choice and Welfare*, 41(4):989–998, 2013.

[10] J.-P. Doignon and J.-C. Falmagne. A polynomial time algorithm for unidimensional unfolding representations. *Journal of Algorithms*, 16(2):218–233, 1994.

[11] E. Elkind, P. Faliszewski, M. Lackner, and S. Obraztsova. The complexity of recognizing incomplete single-crossing preferences. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 865–871, Jan. 2015.

[12] E. Elkind, P. Faliszewski, and A. Slinko. Clone structures in voters' preferences. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 496–513, June 2012.

[13] P. Emerson. The original Borda count and partial voting. *Social Choice and Welfare*, 40(2):352–358, 2013.

[14] G. Erdélyi, M. Lackner, and A. Pfandler. Computational aspects of nearly single-peaked electorates. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 283–289, July 2013.

[15] B. Escoffier, J. Lang, and M. Öztürk. Single-peaked consistency and its complexity. In *Proceedings of the 18th European Conference on Artificial Intelligence*, pages 366–370, July 2008.

[16] P. Faliszewski, E. Hemaspaandra, and L. A. Hemaspaandra. The complexity of manipulative attacks in nearly single-peaked electorates. *Journal of Artificial Intelligence Research*, 207:69–99, 2014.

[17] P. Faliszewski, E. Hemaspaandra, L. A. Hemaspaandra, and J. Rothe. The shield that never was: Societies with single-peaked preferences are more open to manipulation and control. *Information and Computation*, 209:89–107, 2011.

[18] P. Fishburn. *The Theory of Social Choice*, volume 264. Princeton University Press, 1973.

[19] D. Fulkerson and O. Gross. Incidence matrices and interval graphs. *Pacific Journal of Math*, 15(3):835–855, 1965.

[20] E. Hemaspaandra, L. A. Hemaspaandra, and J. Rothe. Exact analysis of Dodgson elections: Lewis Carroll's 1876 voting system is complete for parallel access to NP. *Journal of the ACM*, 44(6):806–825, 1997.

[21] E. Hemaspaandra, H. Spakowski, and J. Vogel. The complexity of Kemeny elections. *Theoretical Computer Science*, 349(3):382–391, 2005.

[22] J. Kemeny. Mathematics without numbers. *Daedalus*, 88:577–591, 1959.

[23] K. Konczak and J. Lang. Voting procedures with incomplete preferences. In *Proceedings of the 1st Multidisciplinary Workshop on Advances in Preference Handling*, pages 124–129, Aug. 2005.

[24] M. Lackner. Incomplete preferences in single-peaked electorates. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 742–748, July 2014.

[25] N. Mattei, J. Forshee, and J. Goldsmith. An empirical study of voting rules and manipulation with large datasets. In *Proceedings (Workshop Notes) of the 4th International Workshop on Computational Social Choice*, Sept. 2012.

[26] N. Mattei and T. Walsh. PrefLib: A library for preferences. In *Proceedings of the 3rd International Conference on Algorithmic Decision Theory*, pages 259–270, Nov. 2013.

[27] J. A. Mirrlees. An exploration in the theory of optimum income taxation. *The Review of Economic Studies*, 38(2):175–208, 1971.

[28] H. Moulin. On strategy-proofness and single peakedness. *Public Choice*, 35(4):437–455, 1980.

[29] N. Narodytska and T. Walsh. The computational impact of partial votes on strategic voting. In *Proceedings of the 21st European Conference on Artificial Intelligence*, pages 657–662, Aug. 2014.

[30] M. Schulze. A new monotonic and clone-independent, reversal symmetric, and Condorcet-consistent single-winner election method. *Social Choice and Welfare*, 36(2):267–303, 2011.

Theory Choice, Theory Change, and Inductive Truth-Conduciveness

[Extended Abstract]

Konstantin Genin
Carnegie Mellon University
konstantin.genin@gmail.com

Kevin T. Kelly
Carnegie Mellon University
kk3n@andrew.cmu.edu

ABSTRACT

Synchronic norms of theory choice—traditional to the philosophy of science—restrict the theories one can choose in light of given information. Ockham's razor is a famous example. How can one argue that these biases are truth-conducive, without begging the question with material assumptions?

Diachronic norms of theory change—as studied in belief revision—restrict how one should change one's current beliefs in light of new information. How do the diachronic norms relate to the synchronic norms? Furthermore, is there some sense in which the diachronic norms are truth-conducive?

If one insists upon an overly strict standard of truth-conduciveness in inductive contexts, the epistemic justification of inductive norms becomes intractable. Theoretical virtues are not guaranteed to *indicate* truth with a low chance of error, the way litmus paper indicates pH. But there is a spectrum of truth-conduciveness concepts, ranging from the strict standard of truth-indicativeness to the weak standard of mere convergence in the limit. Neither extreme suffices for epistemic justification. The former is too strict to apply and the latter, notoriously, mandates no short-run norms at all. There are more nuanced concepts of optimally *direct* convergence to the truth, lying between these extremes. We consider two such concepts: convergence with minimal reversals of opinion and convergence with minimal cycles of opinion.

We show how the rationality principles of belief revision can be thought of as truth-conducive norms of direct convergence. Furthermore, we prove that preferring simple, falsifiable theories is a necessary condition for satisfying the norms of rational theory change. The results forge deep and, perhaps, surprising connections between synchronic rationality norms, diachronic rationality norms, and the truth-conduciveness of both.

Keywords

theory choice, simplicity, Ockham's razor, learning, belief revision, reliability, formal epistemology

1. INTRODUCTION

This work is concerned with three things: the synchronic norms of theory *choice*, the diachronic norms of theory *change*, and the justification of these norms by their *reliability,* or *truth-conduciveness.*

Synchronic norms of theory choice restrict the theories one can choose in light of given, empirical information. They are the traditional purview of the philosophy of science. But how do they facilitate arrival at true theories? Are they better than other means toward that end and, if so, in what sense? For example, it is widely agreed that scientific theory choice proceeds in accordance with a bias toward simpler or more sharply testable theories. However, the truth might not be simple, in which case that bias would probably lead to error. So how can one argue that the characteristic scientific biases toward simplicity or testability are truth-conducive, without begging the question with material assumptions?

Diachronic norms of theory change restrict how one should change one's current beliefs in light of new information. The crucial difference from synchronic norms is dependency upon one's prior beliefs. Such norms are studied propositionally in belief revision theory and non-monotonic logic and quantitatively in Bayesian epistemology. The question of truth-conduciveness arises for such norms, just as it does for norms of theory choice. Furthermore, there is the additional question how diachronic norms relate to the more traditional, synchronic ones.

If one insists upon an overly strict standard of truth-conduciveness in inductive contexts, the crucial question of truth-conduciveness becomes intractable. Theoretical virtues are not guaranteed to *indicate* truth with a low chance of error, the way litmus paper indicates pH—inductive inferences in accordance with the rationality principles are subject to arbitrarily high chances of error, because the available information can probably be arbitrarily similar, regardless of which conclusion is true.

Maturity is a matter of ceasing to demand the impossible. In that spirit, it makes more sense to adjust the standards of truth-conduciveness to the intrinsic difficulty of the inference problem one faces—a view we call *feasibility-contextualism.* Feasibility contextualism presupposes a spectrum of alternative concepts of truth-conduciveness. Just such a spectrum is routinely studied in the subject known as formal learning theory, which studies concepts of truth-conduciveness ranging from the very strict standard of truth-indicativeness to the very weak standard of mere convergence to the truth in the limit. Neither extreme suffices for epistemic justification. The former is too strict to apply without question-begging assumptions and the latter, notoriously, mandates no short-run norms at all, since convergence in the limit is compatible with any inductive behavior whatever in the short run. In between these extremes, however, are more nuanced concepts of optimally *direct* convergence to the truth. We consider two such concepts in this paper: convergence with minimal reversals of opinion and convergence with minimal cycles of opinion. Since a strategy is conducive to a goal insofar as it leads as directly as possible to the goal, we view directness of approach to the truth as *constitutive* of truth-conduciveness and, hence, of epistemic justification, rather than as an auxiliary, "pragmatic" consideration.

In this work, we show how the rationality principles of belief revision can be thought of as truth-conducive norms of maximally direct convergence in the sense just described. Furthermore, we prove that preferring simple, falsifiable theories (Ockham's razor) is a necessary condition for achieving optimally truth-conducive performance. The results forge deep and, perhaps, surprising connections between synchronic rationality norms, diachronic rationality norms, and the truth-conduciveness of both.

1.1 Reliability and the Norms of Theory Choice

It is commonplace to observe that science seeks true theories about the world. But that banal observation raises a profound question about scientific method: how, and in what sense, do such hallmark scientific values as simplicity, precision, scope, and novelty help one find true theories? To demand an answer to that question is to demand an *epistemic justification* of scientific values [16, 24, 12]. One goal of this work is to provide such a justification for Ockham's razor, the pervasive scientific bias in favor of simple theories.

An epistemic justification of Ockham's razor is traditionally understood to be a demonstration that simpler theories are *more likely to be true.*[1] That narrow, synchronic concept of justification makes a hopeless conundrum out of Ockham's razor. It has led theorists into metaphysical speculations less plausible than the scientific conclusions they are meant to justify. Both Kepler and Dirac expressed the conviction that Nature loves mathematical elegance, and that physicists ought to adopt the same passion to more surely uncover her secrets [29, 11]. It has led others to give up on epistemic justification entirely: "no one has shown that any of these rules is more likely to pick out true theories than false ones. It follows that none of these rules is epistemic in character" [25].

There is no shortage of non-epistemic justifications. Predictive accuracy, not truth, is the target of frequentist justifications: a bias toward simple theories prevents over-fitting and improves prediction when extrapolating from small samples [1, 13, 38]. But while they may be more predictively accurate at small sample sizes, simple theories are not more likely to be true in any objective sense of likelihood. Akaike's method does not even converge to the true theory in the limit of infinite data. Some frequentists take that to be a design feature, rather than a flaw, and warn against methods that both impose penalties on complexity *and* converge to the truth [26]. Bayesians, on the other hand, explicate Ockham's razor as the result of conditioning over a wide class of plausible, prior probabilities that impose flattish distributions over theoretical parameters [17, 5, 39, 30]. But that does not begin to explain how such prior probabilities lead one to true theories better than alternative biases would—unless one begs the question by appealing to the prior probabilities themselves. The question of *epistemic* justification, if not begged, is dodged.

The point can be sharpened with a bit of terminology. Say that a method is *truth-indicative* if at every stage of inquiry the theory it selects is probably true. But truth-indicative performance is impossible in inductive inference problems—insisting on achieving that impossible synchronic standard leads to inductive skepticism. More plausibly, one can entertain a range of weaker concepts of *diachronic truth-conduciveness*, and understand epistemic justification as achievement of the strongest performance possible for the problem one faces.[2] Weaker demands do not fall short of epistemic demands. They are, rather, the *appropriate* epistemic demands, in light of the intrinsic difficulty of the task at hand.

Over half a century ago, Carnap already sketched the idea in *On Inductive Logic* [8]. He recognized, that for inductive methods, synchronic truth-indicativeness is *too high* a standard: "the fact that the truth of the predictions reached by induction cannot be guaranteed does not preclude a justification in a weaker sense". On the other hand, Reichenbach's diachronic norm of limiting convergence is *too low* a standard:

> Reichenbach is right in the assertion that any procedure which does not [converge to the truth in the limit] is inferior to his rule of induction. However, his rule, which he calls "the" rule of induction, is far from being the only one possessing the characteristic. The same holds for an infinite number of other rules of induction. ... Therefore we need a more general and stronger method for examining and comparing any two given rules of induction ... [8, p.]

The relevant notions of truth-conduciveness have to lie somewhere between those two extremes. If they are to be feasible, they must relax truth-indicativeness. If they are to mandate interesting short-run methodological principles like Ockham's razor, they must demand more than mere limiting convergence. We propose that such notions can be developed by adapting and refining existing concepts from formal learning theory.

1.2 Learning Theory and Truth-Conduciveness

[1] Baker [3] is representative: "justifying an epistemic principle requires answering an epistemic question: why are parsimonious theories more likely to be true?" The demand is trivial if "likely" is understood subjectively, so we understand it objectively, in the sense of a guaranteed low chance of error.

[2] In statistics, the distinction between truth-indicative and truth-conducive methods is closely tracked by the distinction between uniformly and point-wise consistent methods. Both types of consistency entail convergence to the truth in the limit, but for uniformly consistent methods, the probability and severity of error can be quantified and bounded at each sample size. For point-wise convergent methods no such guarantees can be given. Of course, we *prefer* uniformly convergent methods, but these do not always exist. No statistician claims that using a point-wise convergent method is not epistemically justified when there is no better alternative.

Formal learning theory is a mathematical framework for studying inductive problems and the methods that solve them [33, 15, 31, 18]. As in computational complexity theory, inductive problems are classified by their intrinsic difficulty. Inductive methods are justified so long as they solve a problem as efficiently as problems of comparable complexity can be solved. From the perspective of formal learning theory, it makes no more sense to demand truth-indicative performance in inductive problems than it does to demand general polynomial time solutions to NP-hard problems. The demands of epistemic justification are kept proportionate to epistemic complexity.

The baseline notion of truth-conduciveness in formal learning theory is limiting convergence: methods eventually settle on the truth as information accumulates, without ever becoming certain that the future holds no surprises in store. As Carnap observed, limiting convergence is compatible with any arbitrary behavior in the short-run. To narrow the field of admissible methods, learning theorists have developed several refinements of limiting convergence that fall short of short-run guarantees. One possible refinement is to require methods to minimize the number of *mind changes* on the way to convergence [33, 9, 36, 28, 19, 21, 22]. Another refinement guards against "U-shaped learning," wherein learners conjecture a theory, reject it, and then return to it again [7, 6]. We propose that these convergence criteria—lying midway between truth-indicativeness and mere limiting convergence—can answers Carnap's challenge. If scientists are in pursuit of truth, then virtuous inquiry ought to exhibit the virtues of pursuit. If the target of pursuit is evasive, then a certain amount of swerving may be expected. But false starts and U-turns ought to be avoided if possible—*virtuous pursuit is as direct as the problem situation allows.* We show how optimal truth-conduciveness mandates a preference for simple theories. That solves the traditional puzzle of justification for Ockham's Razor, a hallmark virtue of theory choice.

1.3 Belief Revision and Scientific Inquiry

Belief revision is an alternative, formal framework in which to analyze belief change driven by new information. Reliability and truth-conduciveness are not central to the belief revision framework. Gärdenfors seems indifferent—if not hostile—to concerns about truth:

> [T]he concepts of truth and falsity are *irrelevant* for the analysis of belief systems. These concepts deal with the relation between belief systems and the external world, which I claim is not essential for an analysis of epistemic dynamics. ... My negligence of truth may strike traditional epistemologists as heretical. However, one of my aims is to show that many epistemological problems can be attacked without using the notions of truth and falsity [14, p. 20].

Instead, belief revision theorists derive epistemic justification from conformity with a set of idealized postulates that govern rational belief change. The postulates are usually motivated by considerations of *preservation*, or *minimal change*: the injunction to (1) add only those new beliefs and (2) remove only those old beliefs, that are absolutely compelled by incorporation of new information.[3] It is not obvious that those postulates of rationality have anything to do with truth-conduciveness. In §3.1, we demonstrate that such a connection does, in fact, exist: we show that a weakened version of the rationality postulates is equivalent to a truth-conduciveness norm from formal learning theory, once the requirement of limiting convergence has been imposed.

It is also not obvious that there should be any connection between the rationality postulates that belief revision theorists take to govern theory change, and the synchronic theoretical virtues investigated by philosophers of science. Rott expresses a hope that such connections exist:

> In his joint book with J.S. Ullian, The Web of Belief (1978), Quine has added more virtues that good theories should have: modesty, generality, refutability, and precision. Again, belief revision as studied so far has little to offer to reflect the quest for these intuitive desiderata. Except for the issue of conservatism, Quine's list is one of theory choice rather than theory change in that it lists properties that a good posterior theory should have, independently of the properties of the prior theory. It is a strange coincidence that the philosophy of science has focussed on the monadic (nonrelational) features of theory choice, while philosophical logic has emphasized the dyadic (relational) features of theory change. I believe that it is time for researchers in both fields to overcome this separation and work together on a more comprehensive picture [34, p. 15].

In §5, we demonstrate that theoretical refutability is a necessary condition for the rationality postulates of belief revision, in light of the requirement of limiting convergence. What's more, we show that theoretical refutability is equivalent to a version of theoretical simplicity.

2. PROBLEMS AND SOLUTIONS

We first give a minimal characterization of the context of empirical inquiry. Let W be a set of possible worlds. A ***proposition*** is a set $P \subseteq W$. The set of all propositions is denoted $\mathcal{P}(W)$. The contradictory proposition is $\varnothing$ and the necessary proposition is W. We assume the usual correspondence between logical and set-theoretic operations: $P \wedge Q = P \cap Q$, $P \vee Q = P \cup Q$, $P^c = W \setminus P$ and P entails Q iff $P \subseteq Q$.

Some propositions correspond to possible ***information states***. Propositional information is understood to be true. Examples include propositions concerning discrete experimental outcomes and inexact measurements of continuous quantities. Let $\mathcal{I} \subseteq \mathcal{P}(W)$ be the set of all possible information states one might be in. We denote the set of all information states in world w as: $\mathcal{I}(w) = \{E \in \mathcal{I} : w \in E\}$. For $P \subseteq W$, we let $\mathcal{I}(P) = \bigcup_{w \in P} \mathcal{I}(w)$. The set of information states $\mathcal{I}$ is an ***information basis*** iff the following postulates are satisfied. I1. $\bigcup \mathcal{I} = W$; I2. If $A, B \in \mathcal{I}(w)$,

[3] Rott [34] questions whether the rationality postulates of belief revision are correctly thought of as principles of minimal change. Alternatively, one can think of them as monotonicity principles. That also suggests a rapprochement with the learning-theoretic viewpoint: adherence to the rationality principles throughout the course of inquiry guarantees a certain degree of monotonicity—or "directness" – of convergence to the truth.

then $A \cap B \in \mathcal{I}(w)$.[4] I3. $|\mathcal{I}| \leq \omega$. The second thesis is the most interesting: it ensures that information accumulates through time. The third thesis is partly for mathematical tractability, but it is also well-motivated by Turing's [37] argument that infinite gradations of input information are indistinguishable.

The closure of the information basis under arbitrary union, $\mathcal{I}^*$, determines a topological space. An open set of $\mathcal{I}^*$ is a ***verifiable*** proposition — if true, there is information available that entails it. A closed set of $\mathcal{I}^*$ is ***refutable*** — its complement is verifiable. A set is ***locally closed*** in $\mathcal{I}^*$ iff it is closed in an open subspace of $\mathcal{I}^*$. The locally closed sets are ***verifutable*** propositions — if true, it is eventually verified that they are refutable. A set is ***constructible*** in $\mathcal{I}^*$ iff it is a finite union of locally closed sets. A set is Σ^0_2 in $\mathcal{I}^*$ iff it is a countable union of locally closed sets. The topological ***closure*** of $P \subseteq W$, written as $\overline{P}$, is the set $\{w : \mathcal{I}(w) \subseteq \mathcal{I}(P)\}$ — the set of worlds where P is never refuted. The topological ***frontier***, written as $\overset{\vee}{P}$, is defined as $\overline{P} \setminus P$ — the set of worlds where P is false, but never refuted. The following characterization of locally closed sets will prove useful in what follows.

Theorem 1. *P is locally closed iff $\overset{\vee}{P}$ is closed.*

A ***question*** $\mathcal{Q}$ on W is a countable partition of W into mutually exclusive and exhaustive ***answers***. We denote the unique answer true in $w \in W$ as $\mathcal{Q}(w)$. For $P \subseteq W$, we let $\mathcal{Q}(P) = \bigcup_{w \in P} \mathcal{Q}(w)$. A question is locally closed iff each of its answers is. One question refines another, if each answer of the one entails some answer of the other. An ***empirical problem*** is a triple $\mathfrak{P} = (W, \mathcal{I}, \mathcal{Q})$, specifying the set of possibilities entertained (W), the information *provided* $(\mathcal{I})$ and the information *demanded* $(\mathcal{Q})$. A ***learner*** is a function from information states to conjectures $\lambda : \mathcal{I} \to \mathcal{Q}^*$, where $\mathcal{Q}^*$ is the closure of $\mathcal{Q}$ under arbitrary disjunction. A learner λ ***solves*** $\mathfrak{P}$ ***in the limit*** iff for all $w \in W$, there is $E \in \mathcal{I}(w)$ such that for all $F \in \mathcal{I}(w)$, $\lambda(E \cap F) = \mathcal{Q}(w)$. Call such a learner a ***solution*** to $\mathfrak{P}$. No matter which world is the true one, a solution eventually converges on the answer true in that world. For learner λ and $w \in W$, define $\mathsf{Lock}(\lambda, w)$ as the set of all $E \in \mathcal{I}(w)$ such that if $F \in \mathcal{I}(w)$, then $\lambda(E \cap F) \subseteq \mathcal{Q}(w)$. We call these the ***locking*** propositions, following the usage in [31]. An empirical problem is ***solvable*** iff there is a learner that solves it in the limit. The following theorem characterizes solvable problems.[5]

Theorem 2. *The following propositions are equivalent:*

1. *Problem $\mathfrak{P} = (W, \mathcal{I}, \mathcal{Q})$ is solvable in the limit;*
2. *Each answer $Q \in \mathcal{Q}$ is a Σ^0_2 proposition;*
3. *Question $\mathcal{Q}$ is refined by a locally closed question $\mathcal{Q}'$.*

We have two easy corollaries. Say that a learner is *consistent* iff $\lambda(E) \cap E$ is non-empty, for all non-empty $E \in \mathcal{I}$. Then:

Theorem 3. *Every solvable problem has a consistent solution.*

New information E may be thought of as shifting the given problem $\mathfrak{P}$ to the restricted problem

$$\mathfrak{P}|_E \;=\; (E, \mathcal{I}|_E, \mathcal{Q}|_E),$$

where $\mathcal{I}|_E$ is the set of all information states $F \cap E$ such that F is in $\mathcal{I}$, and $\mathcal{Q}|_E$ is the set of all $H \cap E$ such that H is an answer to $\mathcal{Q}$. Then:

Theorem 4. *If $\mathfrak{P}$ is a solvable problem and $C \subseteq W$, then $\mathfrak{P}|_C$ is a solvable problem.*

3. REFINING LIMITING SOLVABILITY

Solution in the limit furnishes a minimal notion of inductive truth-conduciveness. But as many have observed, it enforces no interesting methodological norms, since it is consistent with any short-run behavior. We introduce some notions of truth-conduciveness that refine solution in the limit. As we will show, these notions of truth-conduciveness are weak enough to be feasible in a broad class of problems, and strong enough to mandate interesting norms of theory choice. We also demonstrate how these norms relate to the diachronic norms of theory change advocated as principles of rationality in belief revision and non-monotonic logic.

We first define three different forms of non-monotonicity, in decreasing generality. A ***reversal sequence*** is a sequence $(A_i)^n_{i=0}$, where each $A_i \in \mathcal{Q}^* \setminus \{\varnothing\}$, and $A_{i+1} \subseteq A^c_i$. A ***cycle sequence*** is a reversal sequence such that $A_n \subseteq A_0$. For reversal sequences a, b of length n, define the *severity* pre-order $b \leq a$ to hold iff $B_i \subseteq A_i$, for each $i \leq n$, where $a \leq b$ means that a reverses as *severely* as b. The strict relation $<$ holds iff $\leq$ holds one way and not the other. For example, $(A, B, C) < (A \cup D, B, C)$. A reversal sequence $a = (A_i)^n_{i=0}$ is ***forcible*** in $\mathfrak{P}$ iff for every λ that solves $\mathfrak{P}$, there is a nested set of information states $e = (E_i)^n_{i=0}$, such that $\lambda(e) = (\lambda(E_i))^n_{i=0} \leq a$. We can characterize the forcible sequences as follows.

Theorem 5. *If $\mathfrak{P}$ is solvable, then reversal sequence $a = (A_0, \ldots, A_n)$ is* forcible *in $\mathfrak{P}$ iff*

$$A_0 \cap \overline{A_1 \cap \overline{\ldots \cap \overline{A_{n-1} \cap \overline{A_n}}}} \neq \varnothing.$$

We define the ***forcible paths in*** $\mathfrak{P}$ by a recursion on the length of paths.

$$\Pi_1 = \{A : A \in \mathcal{Q}\};$$
$$\Pi_n = \{A \cap \overline{B} : A \in \mathcal{Q} \text{ and } B \in \Pi_{n-1}\},$$

and $\mathsf{Path}(\mathfrak{P}) = \bigcup_{i \in \mathbb{N}} \Pi_i$.

3.1 Avoiding Cycles

One refined notion of truth-conduciveness is to avoid cycles altogether. A learner λ is ***cycle free*** iff there exist no nested set of information states e, such that $\lambda(e)$ is a cycle sequence. That truth-conduciveness notion is closely related to several norms of rational theory change. Say that a learner λ satisfies ***conditionalization*** iff $\lambda(E) \cap \mathcal{Q}(E \cap F) \subseteq \lambda(E \cap F)$. A learner λ is ***rationally monotone*** iff

[4] This can be weakened to I'2. If $A, B \in \mathcal{I}(w)$, then there is $C \in \mathcal{I}(w)$ such that $C \subseteq A \cap B$. Nothing essential would be changed, though the statements of the theorems and their proofs would be more cumbersome.

[5] A version of this result is given by de Brecht and Yamamoto [10, Theorem 5], where it is couched in the terms of computable analysis. A similar theorem was proven independently by Kelly [20, Corollary 1] in a first-countable setting. It was arrived at independently by Baltag, Gierasimczuk, and Smets [4, Theorem 8]. It is proven for the Baire space by Kelly [18, Proposition 4.10].

$\lambda(E \cap F) \subseteq \lambda(E) \cap \mathcal{Q}(E \cap F)$ whenever $\lambda(E) \cap \mathcal{Q}(E \cap F) \neq \varnothing$.[6] Many authors have suggested that these principles may be too strong to govern inductive inference [14, 35, 27]. However, both these principles are weakened by the following. A learner λ is ***reversal monotone*** iff $\lambda(E \cap F)$ meets $\lambda(E)$ whenever $\lambda(E) \cap \mathcal{Q}(E \cap F) \neq \varnothing$. We show that for learners, reversal monotonicity is equivalent to avoiding cycles. This demonstrates a tight connection between a norm of truth-conduciveness, and two norms of theory change.

Theorem 6. *If consistent λ solves $\mathfrak{P}$, then λ is cycle free iff λ is reversal monotone.*

Although not all solvable problems have cycle-free solutions, every solvable problems is refined by one that does. This universality principle is a direct consequence of Theorem 3 of Baltag, Gierasimczuk, and Smets [4].

Theorem 7. *If $\mathfrak{P} = (W, \mathcal{I}, \mathcal{Q})$ is a solvable problem, then there is $\mathfrak{P}' = (W, \mathcal{I}, \mathcal{Q}')$ such that $\mathcal{Q}'$ is locally closed, $\mathcal{Q}'$ refines $\mathcal{Q}$, and $\mathfrak{P}'$ is solved by a cycle-free learner.*

3.2 Minimizing Reversals

Although in many cases cycles can be avoided altogether, reversals cannot be avoided in inductive problems. But this does not mean that they cannot be minimized. Say that λ is ***reversal optimal*** for $\mathfrak{P}$ iff λ solves $\mathfrak{P}$ and every λ-reversal sequence is forcible. Not all problems have reversal optimal solutions, but we can characterize the ones that do. For $A \in \mathcal{Q}$, define $A_\perp = \{B \in \mathsf{Path}(\mathfrak{P}) : A \cap \overline{B} = \varnothing\}$.

Theorem 8. *Locally closed $\mathfrak{P}$ has a reversal optimal solution iff for every $A \in \mathcal{Q}$, $A \cap \overline{\cup A_\perp} = \varnothing$.*

The characterization is straightforward: a problem has a reversal-optimal solution iff in each answer A, the set of all paths that are *not* forcible from A, $A_\perp$, are topologically separable from A.

4. SIMPLICITY AND FALSIFIABILITY

If we live in a world where bread will always nourish, all information we ever receive will be consistent with bread ceasing to nourish sometime in the future. But the situation is asymmetrical: if we lived in a world where bread ceases to nourish eventually, we would find out sooner or later. There is a natural order that captures the structure of inductive underdetermination between possibilities. For $X, Y \in \mathcal{P}(W)$ set $X \prec Y$ iff $X \subseteq \overset{\vee}{Y}$, and $X \preceq Y$ iff $X \prec Y$ or $X = Y$. If $X \prec Y$, we say that X *faces the problem of induction* with respect to Y, i.e. X is inconsistent with Y, but any *information* consistent with X is consistent with Y. For all $X, Y \in \mathcal{P}(W)$, we say that X is ***simpler*** than Y if and only if $X \prec Y$. That straightforward relation captures many of our simplicity intuitions. Exactly that relation holds between sets of polynomials of lower and higher degree; between nested statistical models; and between universal generalizations like "all ravens are black" and their negations. The $\prec$ relation is not in general a strict order. However, if $\mathcal{Q}$ is locally closed then $\prec$ determines a strict order over the elements of $\mathcal{Q}$.

Theorem 9. *If question $\mathcal{Q}$ is locally closed, then $\prec$ is transitive on $\mathcal{Q}$.*[7]

It follows immediately that $\prec$ is a strict order on $\mathcal{Q}$, and that $\preceq$ is a partial order on $\mathcal{Q}$. Our notion of simplicity is closely related to Popper's. Popper [32] proposes to define simplicity in terms of the falsifiability relation:

> A statement x is said to 'falsifiable in a higher degree' or 'better testable' than a statement y ... if and only if the class of potential falsifiers of x includes the class of the potential falsifiers of y as a proper subclass' ... The epistemological questions which arise in connection with the concept of simplicity can all be answered if we equate this concept with *degree of falsifiability*.

But Popper's notion and ours are not equivalent. To see that, define the class of potential falsifiers of $X \subseteq W$ as the set of all information states inconsistent with X: $\mathcal{F}(X) = \mathcal{I} \setminus \mathcal{I}(X)$. Then X is *more falsifiable* than Y, if $\mathcal{F}(Y) \subseteq \mathcal{F}(X)$, or, equivalently, $X \subseteq \overline{Y}$. That notion of simplicity has the bizarre defect that every proposition is simpler than its consequences. Our notion of simplicity does not face this difficulty, since $X \prec Y$ iff X is more falsifiable than Y, and X and Y are incompatible.

We can make the make the relation between simplicity and falsifiability even more explicit. For any proposition A, define the set of propositions strictly simpler than A as follows: $A_\prec : \bigcup\{B \prec A : B \subseteq W\}$. Say that A is ***simplest*** iff A is minimal in $\prec$, i.e. $A_\prec = \varnothing$. Then it is easy to show that $\cup A_\prec = \overset{\vee}{A}$, and that $A_\prec = \varnothing$ iff A is closed (falsifiable).[8]

5. THE NORMS OF CHOICE

In this section we give two different methodological principles that are necessary for avoiding cycles, and minimizing reversals, respectively. That furnishes the necessary connection between truth-conduciveness and the norms of theory choice. Say that a learner λ is ***Ockham*** iff for all $E \in \mathcal{I}$, $\lambda(E)_\prec \cap E = \varnothing$, i.e. $\lambda(E)$ is always simplest in E. On this conception, Ockham's razor is content-neutral: it does not say whether the conjecture has to be weak or strong, requiring only that there be nothing strictly simpler compatible with current information. That is equivalent to the requirement that the learner's conjecture is *falsifiable* at every stage of inquiry, which is precisely Popper's requirement of severe testability—an enjoiner for bold conjectures, vulnerable before the tribunal of experience. We show that obeying Ockham's razor is necessary for avoiding cycles.

Theorem 10. *If λ solves $\mathfrak{P}$ and λ is cycle free, then λ is Ockham.*

As an immediate consequence of Theorems 7, and 10, we have the following Theorem.

Theorem 11. *If $\mathfrak{P} = (W, \mathcal{I}, \mathcal{Q})$ is a solvable problem, then there is $\mathfrak{P}' = (W, \mathcal{I}, \mathcal{Q}')$ such that $\mathcal{Q}'$ is locally closed, $\mathcal{Q}'$ refines $\mathcal{Q}$, and $\mathfrak{P}'$ is solved by an Ockham learner.*

We state our second methodological principle as follows. A learner λ is ***patient*** iff for all $E \in \mathcal{I}$ and $Q \subseteq \mathcal{Q}(E)$, there

[6] Conditionalization and rational monotonicity are intended as analogues to principles (K^*7) and (K^*8) of AGM revision respectively and the defeasible inference principles of the same name.

[7] This can be weakened to the requirement that every $Q \in \mathcal{Q}$ is constructible.

[8] Recall that a set is closed iff its frontier is empty.

is $Q' \subseteq \lambda(E)$ such that $Q' \cap \overline{Q} \neq \varnothing$.[9] Although Ockham methods can favor some simplest possibilities over others, a patient method must disjoin all simplest possibilities compatible with current information. Furthermore, if a patient conjecture entails the negation of a particular answer, it is because it concedes a simpler possibility. In particular, if $\lambda(E) \subseteq Q^c$, then $\lambda(E) \cap Q_{\prec} \neq \varnothing$. A patient method always has a simplicity-based *reason* for *dis*believing an answer.

THEOREM 12. *If λ is reversal optimal for $\mathfrak{P} = (W, \mathcal{I}, \mathcal{Q})$, then λ is patient.*

As a consequence of theorem 8, we have the following theorem.

THEOREM 13. *Locally closed $\mathfrak{P}$ has a patient, Ockham solution iff for every $A \in \mathcal{Q}$, $A \cap \overline{\cup A_{\perp}} = \varnothing$.*

6. CONCLUSION

In this work we have proposed two norms of inductive truth-conducivness: the avoidance of theoretical cycles, and the minimization of reversals. We have shown that once the requirement of limiting convergence is imposed, avoiding cycles is equivalent to a weakening of two norms of theory change from belief revision and non-monotonic logic. We have also given a topological characterization of theoretical simplicity, and formulated three related norms of theory choice: falsifiability, Ockham's razor, and patience. We have shown that minimizing reversals requires learners to be patient. We have also shown that avoiding cycles requires learners to be Ockham, and always make falsifiable conjectures. This means that the preservation principles from belief revision and non-monotonic logic, combined with the requirement of limiting convergence, all necessitate a preference for simpler, more falsifiable theories. We take this to demonstrate a surprising connection between the Popperian preference for falsifiable theories, and the principles of rational belief change.

7. ACKNOWLEDGEMENTS

We are indebted to Thomas Icard for suggesting connections between our topological conception of simplicity and related work in the semantics of provability logic, which proved to be very fruitful. We are also indebted to Alexandru Baltag, Nina Gierasimczuk, and Sonja Smets, for comments and discussions, for informing us of related work by Debrecht and Yamamoto [10], and for sharing their related results with us, particularly their sharpening of our Theorem 7. We are indebted for discussions following invited presentations of earlier versions of the preceding results at a number of venues: The University of Amsterdam (the 2011 Beth/Vienna Circle Lecture), the ILLC in Amsterdam (2012), the Sorbonne (2012, 2013), Ludwig-Maximiliens Univerity in Munich (2012), the University of Copenhagen (2012), The Perimeter Institute for Theoretical Physics (2010), the CSLI at Stanford (2012), and the Argentinian Society for Philosophical Analysis (2012).

8. PROOFS AND LEMMAS

PROOF OF THEOREM 1. ⇒: Suppose that $A = O \cap C$ for O, C^c open. Then:

$$\begin{aligned} \overset{\vee}{A} = \overline{O \cap C} \cap (O \cap C)^c &= \overline{O \cap C} \cap O^c \cup \overline{O \cap C} \cap C^c \\ &= \overline{O \cap C} \cap O^c, \end{aligned}$$

which is an intersection of two closed sets and therefore closed. The final equality follows from the fact that $\overline{O \cap C} \cap C^c \subseteq \overline{O} \cap \overline{C} \cap C^c = \overline{O} \cap C \cap C^c = \varnothing$. ⇐: Suppose that $\overset{\vee}{A}$ is closed. Since for every A, $A = \overline{A} \setminus \overset{\vee}{A}$, we have that A is a difference of closed sets, and is, therefore, locally closed. □

PROOF OF THEOREM 2. To see that (1) implies (2), suppose that λ solves $\mathfrak{P}$ in the limit. For each w, choose $E_w \in \mathsf{Lock}(\lambda, w)$. Then $\{E_w : w \in W\}$ is a (countable) cover of W by locking information. Let $F_w = \bigcup\{E \in \mathcal{I} : E \subset E_w \text{ and } \lambda(E) \neq \mathcal{Q}(w)\}$. Then $E_w \setminus F_w$ is locally closed. We claim that $A = \bigcup_{w \in A} E_w \setminus F_w$ for each $A \in \mathcal{Q}$. Let $w \in A$. Then $w \in E_w$, and $w \notin F_w$, since E_w is locking for w. So $w \in E_w \setminus F_w \subseteq \bigcup_{w \in A} E_w \setminus F_w$. Suppose that $v \in \bigcup_{w \in A} E_w \setminus F_w$. Then, for some $w \in A$, $v \in E_w \setminus F_w$. Suppose that $\mathcal{Q}(v) = B \neq A$. Then there is $E_v \in \mathsf{Lock}(\lambda, v)$, and $\lambda(E_v \cap E_w) = B \neq A$. But then $v \in F_w$. Contradiction. We have shown that each $A \in \mathcal{Q}$ is a countable union of locally closed propositions.

To see that (2) implies (3), suppose that $A = \bigcup_{i \in \mathbb{N}} L_i$, where L_i is locally closed, for each $A \in \mathcal{Q}$. It is a standard fact that the constructible propositions are closed under finite union and complementation. Letting $C_i = L_i \setminus \bigcup_{j<i} L_j$ we have that each A is a disjoint union of constructible propositions C_i. Furthermore, by a result from [2], proposition P is constructible iff there is a least integer n such that P admits a decomposition into n disjoint, locally closed sets. Therefore, each C_i is a disjoint union of finitely many locally closed propositions. So A is a countable, disjoint union of locally closed sets.[10]

To see that (3) implies (1), suppose that $\mathcal{Q}$ is locally closed. Enumerate the elements of $\mathcal{Q}$ as $A_1, A_2, \ldots$. By Theorem 1, each A_i can be written canonically as a difference of open sets $(\overset{\vee}{A}_i)^c \setminus \overline{A_i}^c$. Define:

$$\lambda(E) = \begin{cases} \min\limits_{A_i \in \mathcal{Q}} E \subseteq (\overset{\vee}{A}_i)^c \text{ and } E \not\subseteq \overline{A_i}^c & \text{if defined;} \\ \mathcal{Q}(E) & \text{otherwise.} \end{cases}$$

Let $w \in A_i$. For $j < i$, either $w \notin (\overset{\vee}{A}_j)^c$, or $w \in \overline{A_j}^c$. Let:

$$E \;=\; \bigcap_{j<i \text{ and } w \in \overline{A_j}^c} \overline{A_j}^c \cap (\overset{\vee}{A}_i)^c.$$

Let $F \in \mathcal{I}(w)$ such that $F \subseteq E$. Then $F \subseteq E \subseteq (\overset{\vee}{A}_i)^c$ and, since $F \in \mathcal{I}(w)$, we have that $F \not\subseteq \overline{A_i}^c$. Furthermore, since $F \in \mathcal{I}(w)$, it follows that $F \not\subseteq (\overset{\vee}{A}_j)^c$, or $F \subseteq \overline{A_j}^c$, for all $j < i$. So $\lambda(F) = \mathcal{Q}(A_i) = \mathcal{Q}(w)$. So $E \in \mathsf{Lock}(\lambda, w)$, as required. □

[9] We say that a problem is ***stratified*** if for all $Q, Q' \in \mathcal{Q}$, $Q \cap \overline{Q'} \neq \varnothing$ entails $Q \preceq Q'$. Most natural problems satisfy this condition. In stratified problems, patience is equivalent to the requirement that every conjecture is co-initial in the simplicity order over the answers. For a discussion of natural problems see [23].

[10] The proof strategy for this step was suggested by Alexandru Baltag in personal communication.

Proof of theorem 3. Recall the solution given in the proof of theorem 2. Let $E \in \mathcal{I}$ be non-empty. Suppose that $\lambda(E) = \mathcal{Q}(E)$. Then $E \cap \lambda(E) = E \cap \mathcal{Q}(E) = E \neq \varnothing$. Suppose that $\lambda(E) = \mathcal{Q}(E_{\sigma(E)} \setminus F_{\sigma(E)})$. Then $E \subseteq E_{\sigma(E)}$ and $E \nsubseteq F_{\sigma(E)}$, so $E \cap \lambda(E) = E \cap \mathcal{Q}(E_{\sigma(E)} \setminus F_{\sigma(E)}) \neq \varnothing$. □

Proof of theorem 4. The result follows immediately from theorem 2 and the fact that a set is locally closed in the subspace topology $\mathcal{I}|_C$ iff it is the intersection of C with a locally closed set. □

Lemma 1. *For arbitrary $B \subseteq W$,*

$$\begin{aligned} &A_n|_B \cap \overline{A_{n-1}|_B \cap \overline{\ldots \cap \overline{A_1|_B \cap \overline{A_0|_B}}}} \\ &\subseteq\ A_n|_B \cap \overline{A_{n-1} \cap \overline{\ldots \cap \overline{A_1 \cap \overline{A_0}}}}, \end{aligned}$$

where, for convenience, we have taken $A|_B$ as shorthand for $A \cap B$.

Proof of Lemma 1. In the base case $n = 1$. We have that: $A_1|_B \subseteq A_1|_B$. By the induction hypothesis, we have that:

$$\begin{aligned} &A_{n+1}|_B \cap \overline{A_n|_B \cap \overline{A_{n-1}|_B \cap \overline{\ldots \cap \overline{A_1|_B \cap \overline{A_0|_B}}}}} \\ &\subseteq\ A_{n+1}|_B \cap \overline{A_n|_B \cap \overline{A_{n-1} \cap \overline{\ldots \cap \overline{A_1 \cap \overline{A_0}}}}} \\ &\subseteq\ A_{n+1}|_B \cap \overline{A_n \cap \overline{A_{n-1} \cap \overline{\ldots \cap \overline{A_1 \cap \overline{A_0}}}}}, \end{aligned}$$

where in each step we have used the fact that if $B \subseteq C$, then $A \cap \overline{B} \subseteq A \cap \overline{C}$. □

Proof of theorem 5. ⇐: Suppose that λ is a solution to $\mathfrak{P}$, and $A_0 \cap \overline{A_1 \cap \overline{\ldots \cap \overline{A_{n-1} \cap \overline{A_n}}}} \neq \varnothing$. Let:

$$\begin{aligned} w_0 &\in A_0 \cap \overline{A_1 \cap \overline{\ldots \cap \overline{A_{n-1} \cap \overline{A_n}}}}; \\ w_{i+1} &\in \bigcap_{j \leq i} E_j \cap A_{i+1} \cap \overline{A_{i+2} \cap \overline{\ldots \cap \overline{A_{n-1} \cap \overline{A_n}}}}; \end{aligned}$$

where $E_j \in \mathsf{Lock}(\lambda, w_j)$. Letting $e = (\bigcap_{j \leq i} E_j)_{i=0}^n$, we have that $\lambda(e) \leq a$.

⇒: Suppose that $\mathfrak{P}$ is solvable. We proceed by induction on n. Base case: $n = 0$. Suppose that $A_0 = \varnothing$. By theorem 3, there is a solution λ such that $\lambda(E) \neq \varnothing$ for all nonempty $E \in \mathcal{I}$. So $a = (A_0)$ is not forcible. For the inductive case, suppose that $A_0 \cap \overline{A_1 \cap \overline{\ldots \cap \overline{A_n \cap \overline{A_{n+1}}}}} = \varnothing$. Let $C = A_1 \cap \overline{\ldots \cap \overline{A_n \cap \overline{A_{n+1}}}}$. To make the subsequent expressions manageable, write $A|_B$ for $A \cap B$. By Lemma 1,

$$\begin{aligned} A_1|_{\overline{C}^c} \cap \overline{\ldots \cap \overline{A_n|_{\overline{C}^c} \cap \overline{A_{n+1}|_{\overline{C}^c}}}} &\subseteq A_1|_{\overline{C}^c} \cap \overline{\ldots \cap \overline{A_n \cap \overline{A_{n+1}}}} \\ &=\ \overline{C}^c \cap C = \varnothing. \end{aligned}$$

So, by the induction hypothesis, $(A_1, ..., A_{n+1})$ is not forcible in the $\mathfrak{P}|_{\overline{C}^c}$ subproblem. Let λ_2 be a solution that never performs that reversal sequence in $\mathfrak{P}|_{\overline{C}^c}$. Furthermore, let λ_1 be a solution to the $\mathfrak{P}|_{\overline{C}}$ subproblem. The solutions to both subproblems exist by theorem 4. Furthermore, we can assume that solution λ_1 is consistent, by proposition 3. Now, define the method:

$$\lambda^*(E) = \begin{cases} \mathcal{Q}(\lambda_1(E \cap \overline{C})) & \text{if } \overline{C} \cap E \neq \varnothing; \\ \mathcal{Q}(\lambda_2(E)) & \text{otherwise.} \end{cases}$$

Since $\overline{C}$ is refutable, it is possible to focus on the $\mathfrak{P}|_{\overline{C}}$ subproblem, until $\overline{C}$ is refuted. Since, by assumption, $A_0 \cap \overline{C} = \varnothing$, the solution λ_1 never conjectures A_0. Therefore, λ_1 cannot perform the reversal sequence $(A_0, A_1, \ldots, A_{n+1})$. Furthermore, since λ_2 never performs the reversal sequence $(A_1, ..., A_{n+1})$, we have that λ^* solves $\mathfrak{P}$, and never performs the reversal sequence $(A_0, A_1, \ldots, A_{n+1})$. □

Proof of theorem 6. ⇐: Suppose that λ performs a cycle. Then there is a nested information sequence $e = (E_i)_{i=0}^n$ such that $\lambda(E_n) \subseteq \lambda(E_0)$ and m such that $1 < m < n$, and $\lambda(E_m) \subseteq \lambda(E_0)^c$. By consistency $\varnothing \neq \lambda(E_n) \cap E_n \subseteq \lambda(E_n) \cap E_m \subseteq \lambda(E_0) \cap \mathcal{Q}(E_m) = \lambda(E_0) \cap \mathcal{Q}(E_0 \cap E_m)$. But $\lambda(E_0) \cap \lambda(E_0 \cap E_m) = \varnothing$.

⇒: Suppose that λ is not reversal monotone. Then there are $E, F \in \mathcal{I}$ such that $\lambda(E \cap F)$ is disjoint from $\lambda(E)$, although $\lambda(E) \cap \mathcal{Q}(E \cap F) \neq \varnothing$. Let $w \in E \cap F$ such that $\mathcal{Q}(w) \subseteq \lambda(E) \cap \mathcal{Q}(E \cap F)$ and $G \in \mathsf{Lock}(\lambda, w)$. Then $\lambda(E \cap F \cap G) = \mathcal{Q}(w) \subseteq \lambda(E)$ and λ performs a cycle. □

Proof of theorem 7. This is a direct consequence of Proposition 2 and Theorem 17 in [4]. □

Proof of theorem 8. ⇒: Suppose that λ is a solution to $\mathfrak{P}$, and for some $A \in \mathcal{Q}$, there is $w \in A \cap \overline{\cup A_\perp}$. Let $E \in \mathsf{Lock}(\lambda, w)$. Then there is $v \in E$, and $P = A_1 \cap \overline{\ldots \cap \overline{A_{n-1} \cap \overline{A_n}}}$, such that $v \in P$ and $P \in A_\perp$. Let $F \in \mathsf{Lock}(\lambda, v)$. Then since P is forcible in $\mathfrak{P}|_{E \cap F}$ by theorem 5, λ performs the reversal $(A, A_1, \ldots, A_n)$, but by assumption, $A \cap \overline{A_1 \cap \overline{\ldots \cap \overline{A_{n-1} \cap \overline{A_n}}}} = \varnothing$, so by theorem 5, that reversal sequence is not forcible.

⇐: Suppose that $\mathfrak{P}$ is locally closed, and that for every $A \in \mathcal{Q}$, $A \cap \overline{\cup A_\perp} = \varnothing$. Enumerate the elements of $\mathcal{Q}$: $A_1, A_2, \ldots$. Define $\mathsf{Root}(E)$ to be the first element in the enumeration such that $E \subseteq (\overset{\vee}{A})^c \cap \overline{\cup A_\perp}^c$, if it exists, and ω otherwise.

$$\lambda(E) = \begin{cases} \mathsf{Root}(E) & \text{if } \mathsf{Root}(E) \neq \omega, \\ \mathcal{Q}(E) & \text{otherwise.} \end{cases}$$

First we show that λ is a solution to $\mathfrak{P}$. Let $w \in A_i$, for some $A_i \in \mathcal{Q}$. As in the proof of theorem 2, there is $E \in \mathcal{I}(w)$ that refutes all A_j such that $j < i$ and $w \in \overline{A_j}^c$. Furthermore, for $j < i$ such that $w \in \overline{A_j}$, clearly $E \nsubseteq (\overset{\vee}{A_j})^c$, since $E \in \mathcal{I}(w)$. So $E \cap (\overset{\vee}{A_i})^c \cap \overline{\cup A_{i\perp}}^c$ is locking for λ in w. We proceed to show that λ is reversal optimal, by induction on the length of reversal sequences. Base case: $n = 1$. Since λ is a solution, any singleton reversal sequence (A_0) is forcible. For the inductive step, suppose that $\lambda(e) = (\lambda(E_i))_{i=0}^n$ is a reversal sequence. WLOG, each $\lambda(E_i) \in \mathcal{Q}$. By hypothesis $(\lambda(E_i))_{i=1}^n$ is forcible. But by construction $\lambda(E_0) \cap \overline{\lambda(E_1) \cap \overline{\ldots \cap \overline{\lambda(E_n)}}} \neq \varnothing$. So by Theorem 5, $\lambda(e)$ is forcible. □

Proof of theorem 9. Let $A, B, C \in \mathcal{Q}$, $A \prec B$, and $B \prec C$. Thus, $A \subseteq \overline{B} \subseteq \overline{C}$. So it remains only to show that $A \neq C$. Each proposition is disjoint from its frontier, so $B \subseteq (\overset{\vee}{B})^c$. Furthermore, if B locally closed $(\overset{\vee}{B})^c$ is open, by theorem 1. So, since $A \subseteq \overset{\vee}{B}$, $(\overset{\vee}{B})^c$ is an open set containing B and disjoint from A. Therefore $B \nprec A$, and $A \neq C$ as required. □

Proof of theorem 10. Suppose that λ is not Ockham. Then for some $E \in \mathcal{I}$ there is $w \in \lambda(E)_{\prec} \cap E$. Let $F \in \mathsf{Lock}(\lambda, w)$, then $\lambda(E \cap F) = \mathcal{Q}(w)$ is disjoint from $\lambda(E)$, although $\lambda(E) \subseteq \mathcal{Q}(E \cap F)$, since w is in the frontier of $\lambda(E)$. Therefore λ is not reversal monotone. So by Theorem 6, λ is not cycle free. □

Proof of theorem 12. Suppose that λ is not patient. Then for some $E \in \mathcal{I}$ there is $Q \subseteq \mathcal{Q}(E)$ such that for all $Q' \subseteq \lambda(E)$, $Q' \cap \overline{Q} = \varnothing$. Therefore $\lambda(E) \cap \overline{Q} = \varnothing$. Let $w \in E \cap Q$ and $F \in \mathsf{Lock}(\lambda, w)$. Then $\lambda(E \cap F) = Q$ is disjoint from $\lambda(E)$. So $(\lambda(E), \lambda(E \cap F))$ is a λ-reversal sequence, but $\lambda(E) \cap \overline{\lambda(E \cap F)} = \varnothing$. So by Theorem 5, that sequence is not forcible, and λ is not reversal optimal. □

Proof of theorem 13. It suffices to notice that the method constructed in the proof of theorem 8 always conjectures an answer closed in the subspace of current information. □

9. REFERENCES

[1] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[2] J. Allouche. Note on the constructible sets of a topological space. *Annals of the New York Academy of Sciences*, 806(1):1–10, 1996.

[3] A. Baker. Simplicity. In E. N. Zalta., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2013 edition, 2013.

[4] A. Baltag, N. Gierasimczuk, and S. Smets. On the solvability of inductive problems: a study in epistemic topology (forthcoming). In *Proceedings of the fifteenth conference on Theoretical Aspects of Rationality and Knowledge*, 2015.

[5] P. S. Bandyopadhayay, R. J. Boik, and P. Basu. The curve fitting problem: A Bayesian approach. *Philosophy of Science*, pages S264–S272, 1996.

[6] L. Carlucci and J. Case. On the necessity of U-Shaped learning. *Topics in Cognitive Science*, 5(1):56–88, 2013.

[7] L. Carlucci, J. Case, S. Jain, and F. Stephan. Non U-shaped vacillatory and team learning. In *Algorithmic Learning Theory*, pages 241–255. Springer Berlin Heidelberg, 2005.

[8] R. Carnap. On inductive logic. *Philosophy of Science*, 12(2):72, 1945.

[9] J. Case and C. Smith. Comparison of identification criteria for machine inductive inference. *Theoretical Computer Science*, 25(2):193–220, 1983.

[10] M. de Brecht and A. Yamamoto. Interpreting learners as realizers for Σ_2^0-measurable functions. *(Manuscript)*, 2009.

[11] P. Dirac. The relation between mathematics and physics. In *Proc. Roy. Soc. Edinburgh*, volume 59, page 122, 1939.

[12] H. Douglas. Inductive risk and values in science. *Philosophy of Science*, pages 559–579, 2000.

[13] M. Forster and E. Sober. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, 45(1):1–35, 1994.

[14] P. Gärdenfors. *Knowledge in Flux.* MIT Press, Cambridge, Mass, 1988.

[15] E. M. Gold. Language identification in the limit. *Information and control*, 10(5):447–474, 1967.

[16] C. Hempel. Valuation and objectivity in science (1983). reprinted in Fetzer, j. (ed.): The philosophy of Carl G, 2001, 1983.

[17] H. Jeffreys. Theory of probability, 1961.

[18] K. T. Kelly. *The Logic of Reliable Inquiry.* Oxford University Press, 1996.

[19] K. T. Kelly. Justification as truth-finding efficiency: How Ockham's razor works. *Minds and Machines*, 14(4):485–505, 2004.

[20] K. T. Kelly. A topological theory of learning and simplicity. *Manuscript*, 2005.

[21] K. T. Kelly. Ockham's razor, empirical complexity, and truth-finding efficiency. *Theoretical Computer Science*, 383(2):270–289, 2007.

[22] K. T. Kelly. Simplicity, truth and probability. In P. S. Bandyopadhyay and M. Forster, editors, *Handbook of the Philosophy of Science Volume 7: Philosophy of Statistics*. North Holland, 2011.

[23] K. T. Kelly, K. Genin, and H. Lin. Realism, rhetoric, and reliability. *Synthese*, (forthcoming) 2014.

[24] L. Laudan. *Science and Values*, volume 87. Cambridge Univ Press, 1984.

[25] L. Laudan. *The Epistemic, the Cognitive, and the Social*, pages 14–23. 2004.

[26] H. Leeb and B. M. Pötscher. Sparse estimators and the oracle property, or the return of hodges? estimator. *Journal of Econometrics*, 142(1):201–211, 2008.

[27] H. Lin and K. T. Kelly. Propositional reasoning that tracks probabilistic peasoning. *Journal of philosophical logic*, 41(6):957–981, 2012.

[28] W. Luo and O. Schulte. Mind change efficient learning. *Information and Computation*, 204(6):989–1011, 2006.

[29] M. Morrison. *Unifying Scientific Theories: Physical Concepts and Mathematical Structures.* Cambridge University Press, 2007.

[30] W. C. Myrvold. A bayesian account of the virtue of unification. *Philosophy of Science*, 70(2):399–423, 2003.

[31] D. N. Osherson, M. Stob, and S. Weinstein. *Systems that Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists.* The MIT Press, 1986.

[32] K. R. Popper. *The Logic of Scientific Discovery.* London: Hutchinson, 1959.

[33] H. Putnam. Trial and error predicates and the solution to a problem of mostowski. *Journal of Symbolic logic*, pages 49–57, 1965.

[34] H. Rott. Two dogmas of belief revision. *The Journal of Philosophy*, pages 503–522, 2000.

[35] G. Schurz. Abductive belief revision in science. In *Belief Revision Meets Philosophy of Science*, pages 77–104. Springer, 2011.

[36] A. Sharma, F. Stephan, and Y. Ventsov. Generalized notions of mind change complexity. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 96–108. ACM, 1997.

[37] A. M. Turing. On computable numbers, with an application to the entscheidungsproblem. *J. of Math*,

58:345–363, 1936.

[38] V. Vapnik. *Statistical Learning Theory*, volume 2. Wiley New York, 1998.

[39] L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107, 2000.

Do players reason by forward induction in dynamic perfect information games?

[Extended Abstract]

Sujata Ghosh
Indian Statistical Institute
Chennai, India
sujata@isichennai.res.in

Aviad Heifetz
The Open University of Israel
Raanana, Israel
aviadhe@openu.ac.il

Rineke Verbrugge
University of Groningen
Groningen, The Netherlands
rineke@ai.rug.nl

ABSTRACT

We conducted an experiment where participants played a perfect-information game against a computer, which was programmed to deviate often from its backward induction strategy right at the beginning of the game. Participants knew that in each game, the computer was nevertheless optimizing against some belief about the participant's future strategy.

It turned out that in the aggregate, participants were likely to respond in a way which is optimal with respect to their best-rationalization extensive form rationalizability conjecture - namely the conjecture that the computer is after a larger prize than the one it has foregone, even when this necessarily meant that the computer has attributed future irrationality to the participant when the computer made the first move in the game. Thus, it appeared that participants applied forward induction. However, there exist alternative explanations for the choices of most participants; for example, choices could be based on the extent of risk aversion that participants attributed to the computer in the remainder of the game, rather than to the sunk outside option that the computer has already foregone at the beginning of the game. For this reason, the results of the experiment do not yet provide conclusive evidence for Forward Induction reasoning on the part of the participants.

Categories and Subject Descriptors

Applied computing [**Law, Social and Behavioural Sciences**]: Economics

Keywords

game theory, experiment, forward induction

1. INTRODUCTION

Backward Induction (BI) is a canonical approach for solving extensive-form games with perfect information. In generic games with no payoff ties, BI yields the unique subgame perfect equilibrium [13, 22]. Nevertheless, BI embodies a conceptual difficulty: in subgames following a deviation of some player (or players) from their BI strategy, it is not obvious why players should necessarily believe that the deviators will 'return to their senses' and realign their behavior in the subgame with the BI dictum. Because such certainty is absent, BI might itself be suboptimal for players who are skeptical about such re-adherence to rationality. Thus, the epistemic assumptions underpinning BI are those of relentlessly reborn optimism (see e.g. the surveys in [21] and [23]), with all players, under all contingencies, commonly believing in everybody's future rationality, no matter how irrational players' past behavior has already proven: "after all, tomorrow is another day!"

An alternative, more sober approach on the part of a player may be to employ Forward Induction (FI) reasoning, and to try to rationalize her opponent's past behavior in order to assess his future moves. For example, even in a subgame where there exists no strategy of the opponent which is consistent with common knowledge of rationality *and* his past behavior, she may still be able to rationalize his past behavior by attributing to him a strategy which is optimal as against a presumed *suboptimal* strategy of hers. Or, even better, it may sometimes be possible for her to attribute to him a strategy which is optimal with respect to a *rational* strategy of hers, which is, though, in return only optimal as against a suboptimal strategy of *his.* If the player pursues this rationalizing reasoning to the highest extent possible [3] and reacts accordingly, she will end up choosing an Extensive-Form Rationalizable (EFR) strategy [4, 20].

EFR strategies may thus be distinct from BI strategies, as an example by Reny [24] shows (see game 1 in Figure 1). Given this difference, it is therefore a completely non-trivial result that in perfect information games with no relevant ties,[1] there is nevertheless a unique EFR *outcome*, which coincides with the unique BI outcome [4,7,8,14,22]. Only when relevant payoff ties are allowed, an outcome-discrepancy between the two solution concepts may appear. In such cases, the EFR outcomes constitute a subset of the BI outcomes [7, 8, 22], and the inclusion may be strict, as demonstrated by Chen and Micali [8] (see game 3 in Figure 1).

We note here that experimental studies in behavioral economics have shown that the backward induction outcome is often not reached in large centipede games. Instead of immediately taking the 'down' option, people often show partial cooperation, moving right for several moves before eventually choosing 'down' [6, 16, 19]. Nagel and Tang [19] suggest that people sometimes have reason to believe that their opponent could be an altruist who usually cooperates by moving to the right and McKelvey and Palfrey [16] suggest that players may believe that there is some possibility that their opponent has payoffs different from the ones the experimenter tries to induce by the design of the game.

[1] That is, where each player has a strict ranking over all the game-tree leaves following each of her decision nodes.

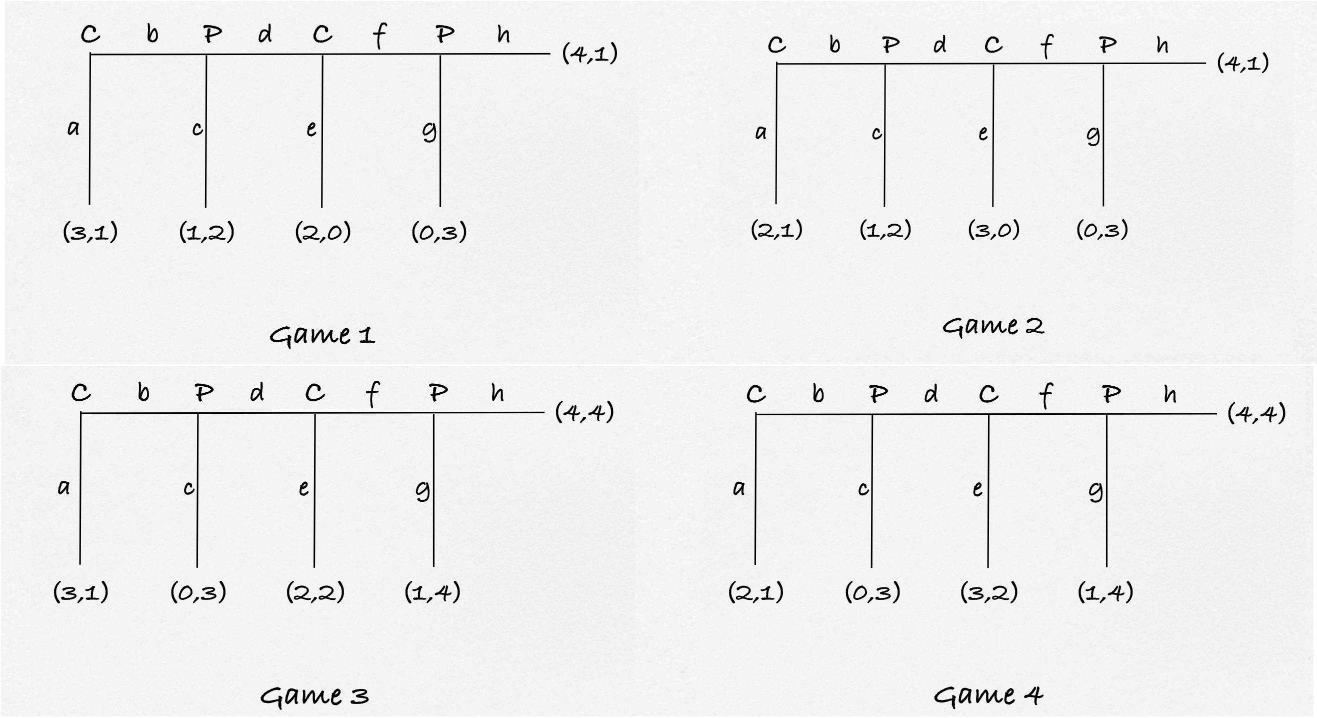


Figure 1: Collection of the main games used in the experiment. The ordered pairs at the leaves represent pay-offs for the computer (C) and the participant (P), respectively.

We could also ask the following question: *Are people inclined to use forward induction when they play a game, and in particular in games like those of Reny or Chen-Micali mentioned above?* This question was the motivation for the experiment on which we report here. Our pivotal interest was to examine participants' behavior following a deviation from BI behavior by their opponent right at the beginning of the game.

1.1 Designing an experiment about forward induction behavior

When designing an experiment to tackle the question whether people are inclined to use FI when they play dynamic perfect-information games, such as the Reny or Chen-Micali games mentioned above, a first challenge was to neutralize repeated-game effects across repetitions of the same game. Such visible repetitions could enable a folk-theorem style augmented cooperation level among participants playing one against the other, bypassing their cooperation opportunities in a one-shot play of the same game (cf. [9]). We chose to address this challenge by letting participants, adult university students with little or no knowledge of game theory, knowingly play against a computer.[2]

We programmed the computer so as to follow, in each repetition of each game, a strategy which is optimal with respect to some strategy of the human participant. This strategy for the computer was decided in advance for each round, so that the computer did not learn from experience in previous games. This information was honestly and simply conveyed to the participants at the beginning of the experiment, by the following item on the instruction sheet (see Appendix A):

> How does the computer reason in each particular game of the experiment?
>
> - The computer thinks that you already have a plan for that game, and it plays the best response to the plan it thinks that you have for that game.
>
> - However, the computer does not learn from previous games and does not take into account your choices during the previous games.

Given that the participants were playing against a computer, a second challenge was to create variability in the appearance of repetitions of the same game, so that each repetition looks different and forces the participant to think anew about her or his strategy in the current repetition. This was achieved by two measures:

- repeating in each round a set of 6 games, distinct in terms of pay-off structures (see more on the games in Section 2); and
- presenting the game to the participant in a different graphical fashion in each round in which the game

[2] Another important advantage of using computer opponents in experiments with dynamic games is that the experimenter can control the strategies used by the computer opponent, which allows better interpretation of the participants' decisions. Using a computer opponent also has disadvantages, for example, players might reason quite differently about their opponent if they know they are playing against a human player. Interestingly, Hedden and Zhang [12] misinformed a part of their subjects that they were playing against a human opponent while in fact everyone was playing against a computer. They found little difference between the decisions of these groups, and only around 10 % of participants expressed a suspicion on an exit questionnaire that they were playing against a computer rather than a person.

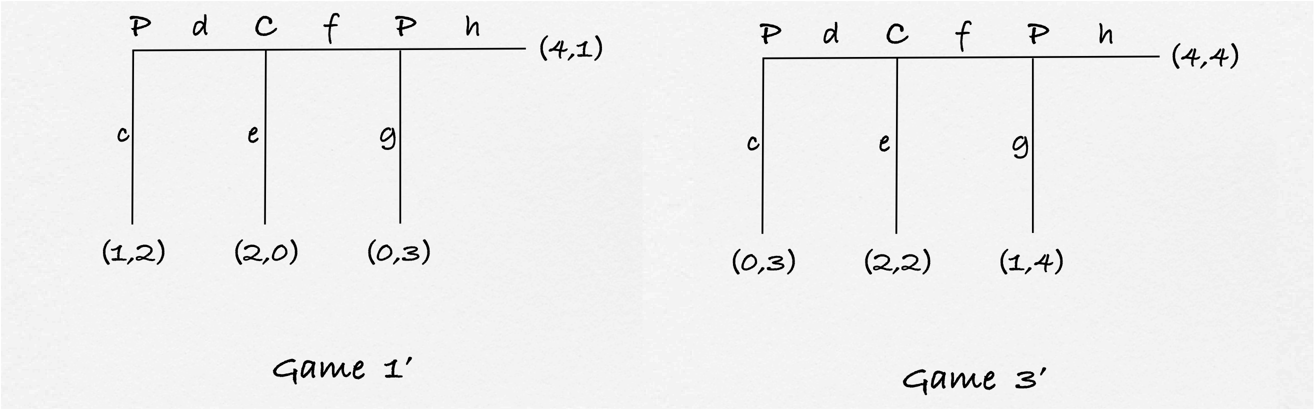


Figure 2: Truncated versions of Game 1 and Game 3. The ordered pairs at the leaves represent pay-offs for C and P, respectively.

was repeated. The game play was animated, and proceeded by consecutive dropping of a marble through trapdoors controlled alternately by the players, leading ultimately to one of several possible bins with orange marbles for the participant and blue marbles for the computer.[3] In repetitions of the same extensive-form game, changes were made in right/left directions of trapdoors in junctions on paths leading to each bin (see, for example, screenshots of different representations of the same game in Appendix C, Figure 6).

In the earlier experiments that investigated FI reasoning in human participants, the experimental games mostly considered an outside option together with some form of imperfect information games, see, e.g., [2, 5, 15, 26]. Such games are more complicated in nature than the dynamical games of perfect information we consider here. The novelty of the current experiment lies in its simplicity, using perfect-information games only.

2. EXPERIMENTAL GAMES

The list of games that were used in the experiment is given in Figure 1 and Figure 2. In these two-player games, the players play alternately. Let C denote the computer, and let P denote the participant. In the first four games (Figure 1), the computer plays first, followed by the participant, and each of the players can play at two decision nodes. In the last two games (Figure 2), which are truncated versions of two of the games in Figure 1, the participant gets first chance to move. We will now discuss the BI and EFR (FI) strategies of all the 6 games; these are summarized in Table 1.

2.1 BI and EFR strategies in four main games

Game 1 has been introduced by Reny [24]. Here, the unique Backward Induction (BI) strategies for player C and player P are $a; e$ and $c; g$, respectively. In case the last decision node of the game is reached, player P will play g (which will give P better payoff at that node) yielding 0 for C. Thus, in the previous node, if reached, C will play e to be better off. Continuing like this from the end to the start of the game (by BI reasoning) it can be inferred that whoever is the current player will play so as to end the game immediately.

Forward induction, in contrast, would proceed as follows. Among the two strategies of player C which are compatible with reaching the first decision node of player P, namely $b; e$ and $b; f$, only the latter is rational for player C. This is because of the fact that $b; e$ is dominated by $a; e$, while $b; f$ is optimal for player C if she believes that player P will play $d; h$ with a high enough probability. Attributing to player C the strategy $b; f$ is thus player P's best way to rationalize player C's choice of b, and in reply, $d; g$ is player P's best response to $b; f$. Thus, the unique Extensive-Form Rationalizable (EFR) strategy of player P is $d; g$, which is distinct from her BI strategy $c; g$. Nevertheless, player C's best response to $d; g$ is $a; e$, which is therefore player C's EFR strategy. Hence the EFR outcome of the game (with the EFR strategies $a; e$, $d; g$) is identical to the BI outcome. This is an instance of the general theorem [1, 4, 8, 14] mentioned in the Introduction, by which in perfect-information games with no relevant payoff ties, the unique BI outcome coincides with the unique EFR outcome (even when, as in this game, for some player the EFR strategy is different from the BI strategy).

Game 2 is popularly known as the Centipede game [25]. Here, the structure of the tree is as in the Reny game (cf. Figure 1, game 1), but the payoffs of player C following a and e are interchanged. As in game 1, the unique Backward Induction (BI) strategies of player C and player P are also $a; e$ and $c; g$, respectively (and the BI outcome is the leaf following a). However, when considering FI reasoning for game 2, unlike in game 1, there does exist a belief of player C with respect to which $b; e$ is optimal. This is the belief that player P is playing with high probability the strategy $d; g$, a strategy that is actually optimal for player P if P believes that C is playing with high probability $b; f$, which in turn is optimal for player C if C believes that player P is playing with high probability $d; h$ and is thus irrational only at the last decision node. For this reason, in game 2 it turns out that $a; e$ and $c; g$ are the unique EFR strategies of the corresponding players, and hence coincide with their unique BI strategies.

Game 3 has been introduced by Chen and Micali [8]. Note that player P has identical payoffs at both leaves following her second and final decision node. As a result, there are two ways to fold the game backwards, and every action of every player at each decision node is a Backward Induction (BI) choice. Consequently, all possible outcomes of the game

[3] The game presentation was inspired by Ben Meijering's 'marble drop' games, also used in [10, 17, 18].

are BI outcomes. However, the strategy $b; e$ of player C is dominated by its strategy $a; e$; and thus $b; e$ is not an Extensive-Form Rationalizable (EFR) strategy for player C. In contrast, $b; f$ is optimal for player C under the belief that player P will pursue $d; h$ with a high enough probability. Hence, if player P finds herself in her first decision node, her best way to rationalize player C's first move of b is to attribute to it the strategy $b; f$, to which only $d; g$ and $d; h$ are best replies. Therefore, the path $b; c$ with the eventual payoffs (0, 3) for players C and P, respectively, is not an EFR outcome of the game. This is an instance of the general result by [7, 8] mentioned in the Introduction, by which the set of EFR outcomes is a (possibly proper) subset of the set of BI outcomes.

Finally, in game 4, the structure of the tree is as in the Chen and Micali game (cf. Figure 1, game 3), but the payoffs of player C following a and e are interchanged with respect to game 3. Here too, every action of every player at each decision node is a BI choice, and hence all possible outcomes of the game are BI outcomes. However, in this case, each of the three strategies that player C has – namely strategy a, strategy $b; e$ and strategy $b; f$ – is a best reply to some conjecture about player P's strategy. Similarly to the game 2 scenario, $b; e$ is a best reply to the conjecture that player P is likely playing $d; g$, and $b; f$ is a best reply to the conjecture that player P is likely playing $d; h$. Thus, for player P, in case her first decision node is reached, both her choices c and d constitute rationalizable (EFR) choices. Hence, in this case, all possible outcomes are EFR outcomes as well, identical to the BI outcomes.

2.2 BI and EFR strategies in truncated games

Game 1′ is a truncated version of game 1, with player P being the starting player. The BI strategy for player P in this game is to play c. In case player P plays d and the first decision node of player C is reached, both BI and EFR strategies for player C are the same − to play e. Thus the EFR strategy for player P in this game is to play c, and the BI and EFR outcomes coincide, as they should in finite perfect-information games without relevant ties.

Game 3′ is a truncated version of game 3, with player P starting the game. Player P has identical payoffs at the leaves following her second decision node. Here again, every action of every player at each decision node is a BI choice, and hence all possible outcomes of the game are BI outcomes. In case the first decision node of player C is reached, each of the two strategies that player C has – namely strategy e and strategy f – is a best reply to some conjecture about player P's strategy. Strategy e is a best reply to the conjecture that player P is likely playing $d; g$, and f is a best reply to the conjecture that player P is likely playing $d; h$. Thus, for player C, in case its second decision node is reached, both its choices e and f constitute rationalizable (EFR) choices. Hence, in this case also, all possible outcomes are EFR outcomes as well, identical to the BI outcomes, in contrast to what happens in game 3.

3. EXPERIMENTAL PROCEDURE

The experiment was conducted at the Institute of Artificial Intelligence (ALICE) at the University of Groningen, The Netherlands. A group of 50 Bachelor's and Master's students from different disciplines at the university took part in this experiment. The participants had little or no knowledge of game theory, so as to ensure that neither backward induction nor forward induction reasoning was already known to them. The participants played the finite perfect-information games in a graphical interface on the computer screen (cf. Figure 3). In each case, the opponent was the computer, which had been programmed to play according to plans that were best responses to some plan of the participant. The participants were instructed accordingly. In each game, a marble was about to drop, and both the participant and the computer determined its path by controlling the orange and the blue trapdoors: The participant controlled the orange trapdoors, and the computer controlled the blue trapdoors. The participant's goal was that the marble should drop into the bin with as many orange marbles as possible. The computer's goal was that the marble should drop into the bin with as many blue marbles as possible.

Games | Strategies	BI strategy	EFR strategy
Game 1	C: $a; e$ P: $c; g$	C: $a; e$ P: $d; g$
Game 2	C: $a; e$ P: $c; g$	C: $a; e$ P: $c; g$
Game 3	C: $a; e, b; e, a; f, b; f$ P: $c; g, d; g, c; h, d; h$	C: $a; e, a; f, b; f$ P: $d; g, d; h$
Game 4	C: $a; e, b; e, a; f, b; f$ P: $c; g, d; g, c; h, d; h$	C: $a; e, b; e, a; f, b; f$ P: $c; g, d; g, c; h, d; h$
Game 1′	C: e P: $c; g$	C: e P: $c; g$
Game 3′	C: e, f P: $c; g, d; g, c; h, d; h$	C: e, f P: $c; g, d; g, c; h, d; h$

Table 1: BI and EFR (FI) strategies for the 6 experimental games in Figures 1 and 2

At first, 14 practice games were played (see Figure 5, Appendix C), which were simpler than the 6 games outlined in Section 2. At the end of each practice game, the participant could see how many marbles he or she had gained in that game, and also the total number of marbles gained so far. These games were presented in increasing levels of difficulty in terms of the reasoning the participants needed to perform with respect to their and the opponent's (computer's) choices, to maximize their gains.

The 14 practice games were followed by 48 experimental games and the participants got access to similar information regarding the number of marbles gained. There were 8 rounds, each comprised of the 6 games that were described in Section 2. Different graphical representations of the same game were used in different rounds (cf. Figure 6, Appendix C). A break of 5 minutes was given after the participants finished playing 4 rounds of the experimental games. The participants earned between 10 and 15 euros for participating in the experiment. The amount depended on the number of marbles won during the experimental phase, and they were told about this before the start of the experiment. They earned 10 euros for participation, and each marble a participant won added 4 cents to the amount. The final amount was rounded off to the nearest 5 cents mark.

At some points during the experimental phase, the participants were asked a multiple-choice question as follows: "When you made your initial choice, what did you think the computer was about to do next?" (cf. Figure 4). Three options were given regarding the likely choice of the com-

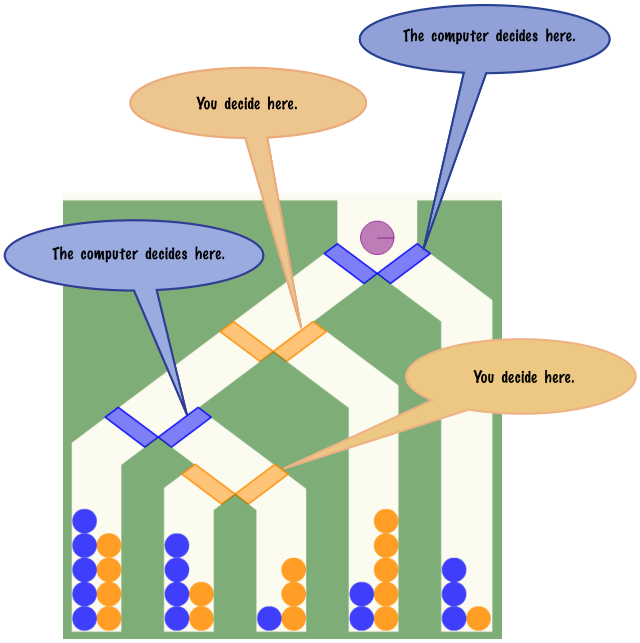


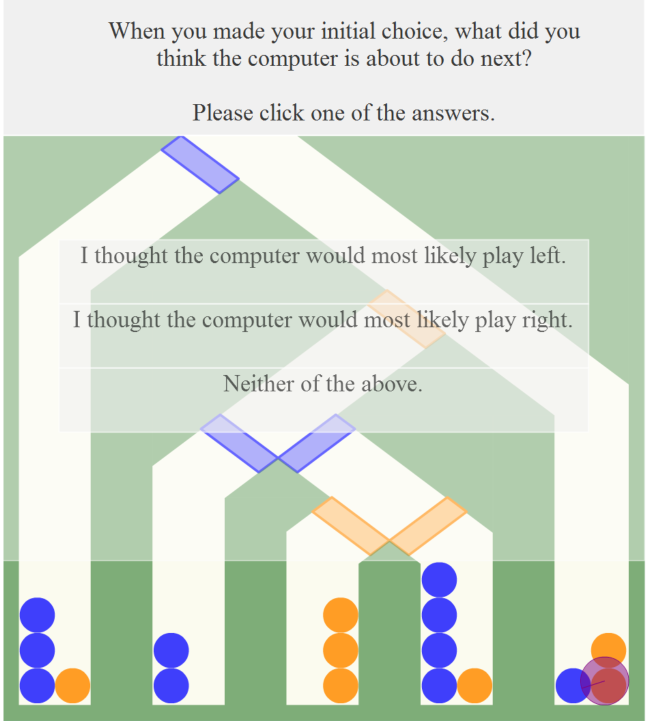
Figure 3: Graphical interface for the participants. The computer controls the blue trapdoors and acquires pay-offs in the form of blue marbles (represented as dark grey in a black and white print), while the participant controls the orange trapdoors and acquires pay-offs in the form of orange marbles (light grey in a black and white print).

puter: “I thought the computer would most likely play left” or “I thought the computer would most likely play right” or “neither of the above”. The first two answers translated to the moves e or f of the computer, respectively. In case of the third answer, we assumed that the participant was undecided regarding the computer’s next choice. The participants had been randomly divided into two groups: Group A and Group B, each consisting of 25 persons. The members of group A were asked the question about the computer’s next possible move once they had played at their first decision node in each game in rounds 3, 4, 7, and 8, whereas the members of group B were asked the same question but less often, namely only in each game in rounds 7 and 8.

At the end of the experiment, each participant was asked the following question: “When you made your choices in these games, what did you think about the ways the computer would move when it was about to play next?” The participant needed to describe the plan he or she thought was followed by the computer on its next move after the participant’s initial choice, in his or her own words. In summary, during the experiment, the participants had to perform the tasks specified in Table 2 in the order given there.

During the 48 games of the experimental phase, played by each participant, a varied amount of data were collected. In particular, for each participant, for each game, for each round of the game, we collected the following data:

- Participant’s decision at his/her first decision node, if the node was reached. In particular, whether move c or d had been played.[4]



When you made your initial choice, what did you think the computer is about to do next?

Please click one of the answers.

I thought the computer would most likely play left.

I thought the computer would most likely play right.

Neither of the above.

Figure 4: Question on computer’s behavior

Step 1	Introduction and instructions.
Step 2	Practice Phase: 14 games.
Step 3	- Experimental Phase: 48 game items, divided into 8 rounds of 6 different games each, in terms of isomorphism class of pay-off structures; - Each of the 6 games occurs once in each round; these games occur in the same order in each round; - Question on computer’s behavior (cf. Figure 4) in several rounds: Group A in rounds 3, 4, 7, 8; Group B in rounds 7, 8.
Step 4	Final Question.

Table 2: Steps of the experiment

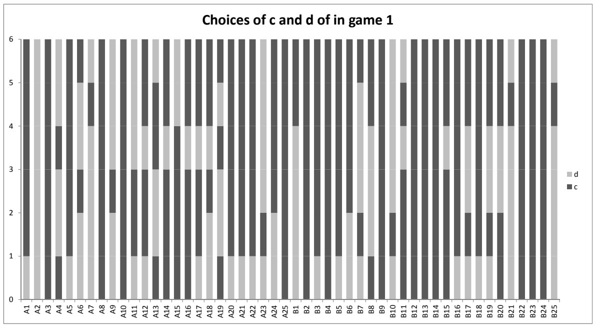
4. RESULTS AND ANALYSIS

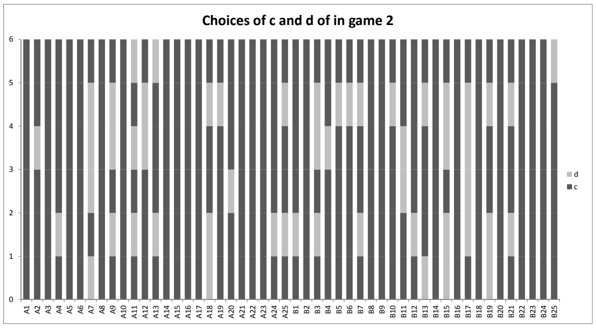
As mentioned above, we report and analyze only the behavior of the participants in their first decision node, that is their choice between actions c or d whenever that decision node was reached. We found no significant variation (Proportion test, p = 0.21) between the behavior of the 25 participants of Group A, who were asked questions (cf. Figure 4) after each game in rounds 3, 4, 7, and 8, and the 25 participants of Group B, who were asked those questions only

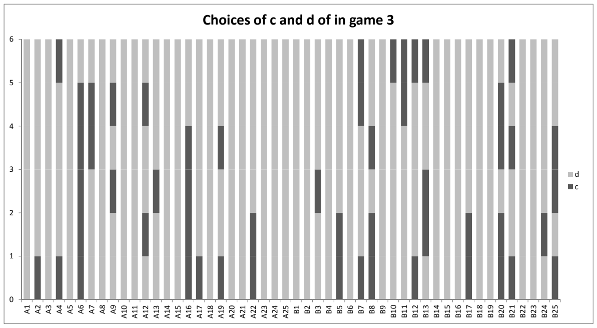
[4] In addition, we also took note of other aspects, such as the participant’s behavior at the second decision node and time taken by the participant at various stages. We leave out the details, because these are not relevant for our main research question, whether participants are applying forward induction. See Appendix B for recorded data types, and see [11] for a typology of players’s reasoning strategies based on these richer data.

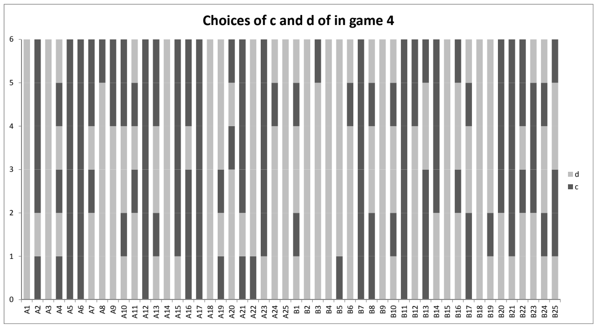
after the games in rounds 7 and 8. Therefore henceforth, we will analyze the data of all 50 participants together.

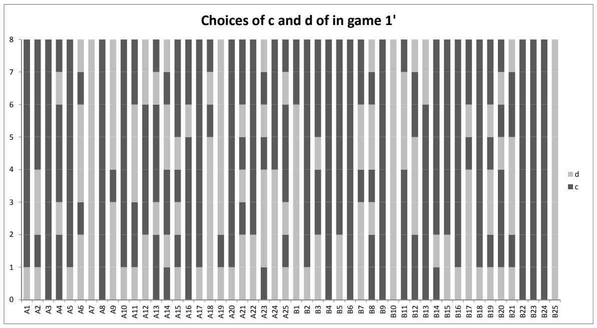
The 6 graphs on the next page give the sequence of choices (across the repetitions of each game) at the first decision node, per participant (named A1 . . . A25, B1 . . . B25). The *dark grey* color corresponds to the rounds the participant played the move *c*, and the *light grey* color corresponds to the rounds the participant played move *d*, whenever the participant's first decision node was reached. They clearly show that *d* was played more often in game 1 than in game 2 (which has the same payoffs as game 1 except for *C*'s payoffs interchanged at two leaves). Moreover, *d* was played more often in game 3 than in game 4 (which similarly has the same payoffs as game 3 except for *C*'s payoffs interchanged at two leaves). These observations may initially suggest corroboration of FI reasoning because (as the reader can check in Table 1), *d* is *P*'s only EFR move in game 1 while *c* is the only EFR move in game 2, and *d* is the only EFR move in game 3 while both *c* and *d* are EFR moves in game 4. This would provide a positive answer to our research question whether players apply forward induction when playing against a computer which sometimes deviates from rational behavior.

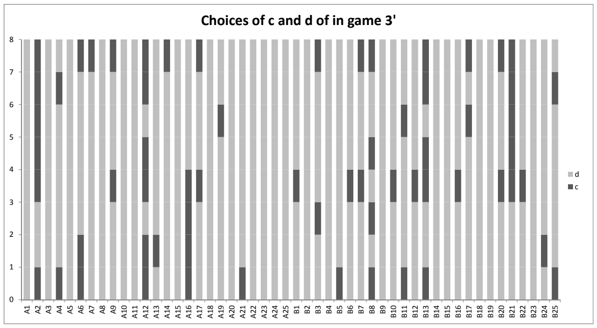












However, a closer look at individual choices while also taking the truncated games $1'$ and $3'$ of Figure 2 into account, casts doubt that these findings can be attributed to any substantial FI reasoning. When comparing game 1 to game $1'$, EFR prescribes *d* in game 1 and *c* in game $1'$ (see Table 1). However, only two participants out of 50 (4%) played *d* much more often in game 1 than in game $1'$;[5] four additional participants (8%) played *d* in game 1 only slightly

[5] The verbal elaboration of one of the two participants at the end of the experiment is indeed compatible with EFR, see Appendix D.

more often than in game $1'$; but 24 other participants (48%) actually played d more often in $1'$ than in 1!

Similarly, when comparing game 3 to game $3'$, EFR prescribes d in game 3, while in game $3'$, both c and d are compatible with EFR. However, only two participants out of 50 (4%) played d much more often in game 3 than in game $3'$; ten additional participants (20%) played d in game 3 only slightly more often than in game $3'$; but 17 other participants (34%) actually played d more often in game $3'$ than in game 3. In summary, comparing games 1 and 3 to their truncated versions does not lend support for FI reasoning.

Now, comparing game 3 to game 4, we find that 47 participants (94%) played d at least as often in game 3 as in game 4, and the remaining three players (6%) played d only slightly less often in game 3 than in game 4. As mentioned at the beginning of this section, at first glance this may then suggest support for EFR behavior (since EFR prescribes d in game 3 and allows for both c and d in game 4). However, because we did not see support for EFR when comparing game 3 to game $3'$, it could very well be that a cardinal effect rather than an ordinal effect has played a role here:

- In game 4, a participant's playing d implied that the computer would have to choose between a payoff 3 that it could reach for certain by going down, and a 'lottery' between the payoffs 1 and 4 that it would meet if it would continue to the right to the next P-node, due to the fact that at P's last decision node, the participant P gains the same payoff of 4 points after either choice. Consequently, most participants might have feared that the computer would go for the certain payoff 3, so preempted that by choosing c.

- In game 3, in contrast, a participant's playing d implied that the computer would have to choose between the relatively low payoff 2 that it could achieve for sure by going down, and again a 'lottery' between the payoffs 1 and 4. Consequently, most participants may have been betting that the computer would go for the 'lottery', and hence chose d.[6]

Similarly, comparing game 1 to game 2, we find that 42 out of 50 participants (84%) played d at least as often in game 1 as in game 2. Here again, at first glance this may seem to lend support for EFR behavior, since EFR prescribes d in game 1 and c in game 2. However, here too, a cardinal effect may have played a role, as follows:

- In game 2, a payoff of 3 may have seemed (in the eyes of most participants) to tempt the computer to go down at its second decision point and settle for it for sure, rather than hoping that the participant would err at the end by choosing h – an error which would yield only a slightly better payoff of 4 to the computer, while C's pay-off would be only 0 if the participant did not err at the end.

- In game 1, in contrast, participants may have attributed a greater temptation to the computer to gamble for the payoff 4 (which is what the computer would get if the participant were to err by choosing h) versus 1 (if the participant did not err); the participant would compare this 'lottery' with what C could settle for with certainty by going down at its second decision point, which is only 2.

These considerations may have led most players to choose d more often in game 1 than in game 2, irrespective of any FI considerations.

5. CONCLUSION

To the best of our knowledge, the experiment carried out and reported here is the first experiment that has been designed to test Forward Induction (FI) behavior (particularly, Extensive-Form Rationalizable (EFR) behavior) in extensive-form games with **perfect information**.

In the experiment, 50 participants played against a computer, which they knew to have been programmed so as not to make deductions or learn from previous game rounds, but rather to optimize, in each round, against some belief about the participant's strategy. Moreover, different rounds of the same game were interspersed in between different rounds of other games, and in different rounds of the same game the game tree was presented to the participants in distinct interactive "marble-drop" forms. Thus, unlike in other experiments where each pair of participants plays repeatedly many rounds of the same game, our design was structured so as to neutralize, as much as possible, repeated-game cooperation considerations on the part of each participant.

In the aggregate, the participants were more likely to respond in a way which is optimal with respect to their best-rationalization EFR conjecture - namely the conjecture that the computer is after a larger prize than the one it has foregone, even when this necessarily meant that the computer has attributed future irrationality to the participant when the computer made the first move in the game. Thus, it appeared that participants did apply forward induction.

However, there exist alternative explanations for the choices of most participants, and such alternative explanations also emerge from several of the participants' free-text verbal descriptions of their considerations (cf. Appendix D), as solicited from them at the end of the experiment. These alternative considerations have to do with the extent of risk aversion that participants attributed to the computer in the remainder of the game, rather than to the sunk outside option that the computer has already foregone at the beginning of the game. For this reason, the results of the experiment do not yet provide conclusive evidence for Forward Induction reasoning on the part of the participants.

In current ongoing work, we are using data from this experiment, such as response times and answers to questions, in order to investigate how participants can be divided into meaningful classes according to other cognitive considerations, for example, whether they are applying quick, instinctive thinking or contemplative, slower deliberation, whether they are applying higher orders of theory of mind, and so on, see [11]. In future work, we aim to investigate which strategies participants actually apply in dynamic games with perfect information in which the opponent occasionally deviates from backward induction. We plan to use new games with different pay-off structures and will perform an eye-tracking study to check the points in the games to which participants attend while reasoning.

[6] Some verbal comments at the end attributed to the computer a 50%-50% belief in this lottery and expected payoff maximization, which is indeed consistent with choosing c in game 4 and d in game 3.

Acknowledgements

We would like to acknowledge research assistants Sourit Manna, Damian Podareanu, and Michiel van de Steeg, without whose help this work would not have been possible. We would also like to thank Harmen de Weerd for his contribution to the graphs in this article. Furthermore, we are grateful to Shaul Tsionit for his statistical advice. Sujata Ghosh and Rineke Verbrugge acknowledge the Netherlands Organisation of Scientific Research for Vici grant NWO 227-80-001 awarded to the latter.

Appendix A: Instruction sheet

- In this task, you will be playing two-player games. The computer is the other player.

- In each game, a marble is about to drop, and both you and the computer determine its path by controlling the orange and the blue trapdoors.

- You control the orange trapdoors, and the computer controls the blue trapdoors.

- Your goal is that the marble drops into the bin with as many orange marbles as possible. The computer's goal is that the marble drops into the bin with as many blue marbles as possible.

- Click on the left trapdoor if you want the marble to go left, and on the right trapdoor if you want the marble to go right.

- How does the computer reason in each particular game?

  > The computer thinks that you already have a plan for that game, and it plays the best response to the plan it thinks that you have for that game.
  >
  > However, the computer does not learn from previous games and does not take into account your choices during the previous games.

- The first 14 games are practice games. At the end of each practice game, you will see how many marbles you gained in that game, and also the total number of marbles you have gained so far.

- The practice games are followed by 48 experiment games. At the beginning of the experiment games, the total number of marbles won will be set at 0 again. At the end of each experiment game, you will see how many marbles you gained in that game, and also the total number of marbles you have gained so far.

- You will be able to start each game by clicking on the "START GAME" button, and move to the next game by clicking on the "NEXT" button.

- At some points during the experiment phase, you will be asked a few questions regarding what guided your choices.

- There will be a break of 5 minutes once you finish 24 of the 48 experiment games.

- The money you will earn is between 10 and 15 euros and depends on how many marbles you have won during the experiment phase. You will get 10 euros for participation, and each marble you win will add 4 cents to your amount. The final amount will be given to you rounded off to the nearest 5 cents mark.[7]

Appendix B: Recorded data types

As mentioned earlier, 50 students participated in this experiment. The participants were first requested to provide the following information:

> Name; Age; Gender; Field of study.

Then they were given instruction sheets mentioning what they were supposed to do (see Appendix A) together with a representative figure (cf. Figure 3) of the graphical interface of the games they were supposed to play. Once they got accustomed with what they were expected to do, the participants played the first 14 practice games. As mentioned in Section 3, at the end of each game, a participant could see how many orange marbles he or she had won till that moment - this was to show how his/her winnings were getting calculated. At the end of the practice phase, the experimental phase began.

Here, each participant played 48 experimental games, playing each of the six games depicted in Figures 1 and 2, eight times, in different representations. During these 48 games, played by each participant, a varied amount of data were collected. For each participant, for each game, for each round of the game, we collected the following data:

- participant's decision at his/her first decision node, if the node was reached. In particular, whether move c or d had been played;

- participant's decision at his/her second decision node, if the node was reached. In particular, whether move g or h had been played;

- time taken by the participant in starting the game, i.e. the time between the moment the game was shown to the participant, and the moment he/she clicked on the "start" button;

- time taken by the participant in making his/her decision at the first decision node, if the node was reached, i.e. the time between the moment the computer passed the playing marble to the participant on its first decision node, and the moment he/she clicked on the next trapdoor for the marble to be dropped;

- time taken by the participant in making his/her decision at the second decision node, if the node was reached, i.e. the time between the moment the computer passed the playing marble to the participant on its second decision node, and the moment he/she clicked on the next trapdoor for the marble to be dropped.

[7] We chose the relatively large 'show-up fee' because Dutch student participants tend to complain in case of large differences in pay between participants. However, most participants attained a fairly large award, so in future we aim to incentivise participants more by offering a lower show-up fee and a higher fee per marble.

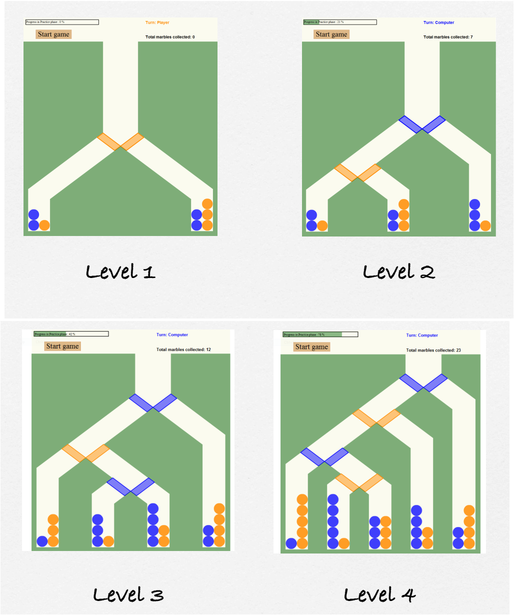


Figure 5: Levels of practice games

The first two items correspond to categorical or qualitative data, whereas the next three, which are response times recorded in milliseconds, correspond to numerical or quantitative data. As mentioned in Section 3, the participants were randomly divided into two groups. namely Group A and Group B, where members of group A were asked to answer a question (cf. Section 3) in rounds 3, 4, 7, 8, and members of group B were asked to answer the same questions but only in rounds 7 and 8. For each participant, depending on the group (Group A or Group B), we collected the following data:

- participant's answer to the given question (cf. Figure 4) at the ends of the rounds in which it was asked. In particular, whether the answer was *e* or *f* or undecided;
- time taken by the participant in giving the answer, i.e. the time between the moment the question appeared on the screen and the moment he/she clicked on his/her choice of answer.

The first data item is categorical, whereas the second one, recorded in milliseconds, is numerical. Finally, at the end of the experiment each participant was asked a final question (cf. Section 3), the answers to which were recorded in a separate sheet. A limited amount of space was given in which the answer was to be formulated.

Appendix C: Experimental interface

During the training phase, the participants were given 14 training games of increasingly difficult levels in terms of number of decision points, as explained in Section 3. Figure 5 shows example games for each of the four levels.

In each of the 8 rounds of the experimental phase, participants were confronted with all 6 games described in Section 2. Different graphical representations of the same game were used in different rounds. As an example, Figure 6 shows six visually different variations of game 1.

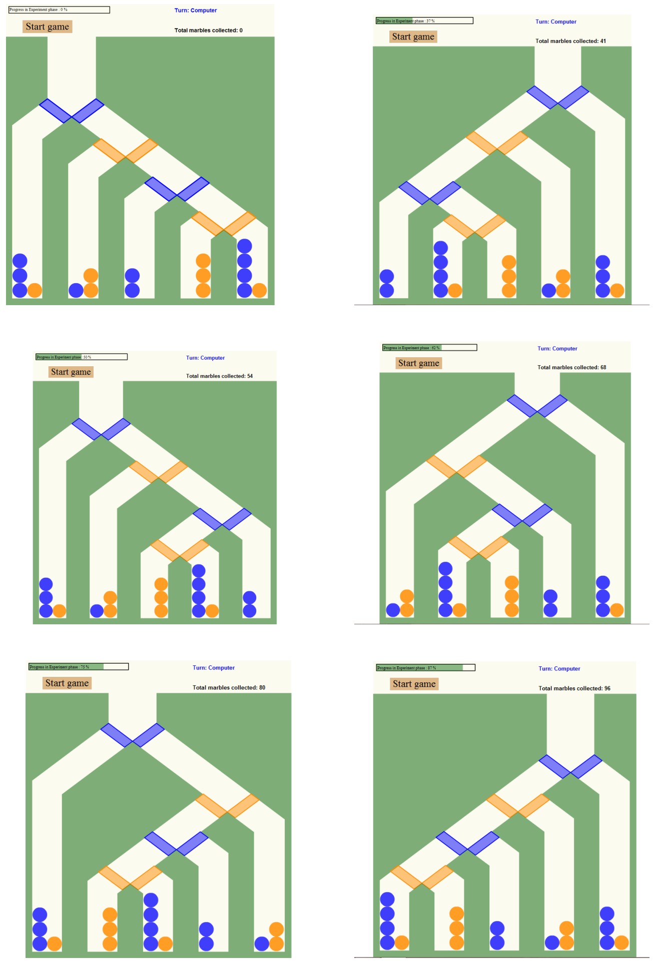

Start game
Turn: Computer
Total marbles collected: 0
Start game
Turn: Computer
Total marbles collected: 41
Start game
Turn: Computer
Total marbles collected: 54
Start game
Turn: Computer
Total marbles collected: 68
Start game
Turn: Computer
Total marbles collected: 80
Start game
Turn: Computer
Total marbles collected: 96

Figure 6: Experimental games, various representations of game 1 of Figure 1

Appendix D: Answers to the final question

As mentioned in Section 3, at the end of the experiment, each participant was asked the following question: "When you made your choices in these games, what did you think about the ways the computer would move when it was about to play next?" The participant needed to describe the plan he or she thought was followed by the computer on its next move after the participant's initial choice, in his or her own words.

We found that one student who had made choices in the game that were consistent with FI reasoning, also provided an answer that suggested FI reasoning:

- "I first thought it would try to maximize the outcomes, taking into account that I would do the same. But I noticed that it did not always do that. Sometimes it did and sometimes it didn't. So after the break, I tried to maximize my outcomes, assuming the computer did the same, but if I noticed that the computer was not assuming that I would maximize my outcomes, I took a risk and I won a lot more."

Here follows a selection of answers provided by the other participants, which shows that participants might have given more importance to risk aversion and/or expected gains, rather than considering the outside option which the computer has already foregone.

- "I thought the computer took the option with the highest expected value. So if on one side you had a 4 blue + 1 blue marble and on the other side 2 blue marbles he would take the option 4+1= 2.5."

- "It was going to take the turn with the highest reward, considering the risk. For example, when the computer can take a reward of 2 marbles instantly or choose to let the ball roll to an orange gate which has the potential of rewarding 4 marbles, the computer would go for the orange gate. With a difference of 1 marble between choices the computer is most likely to take the easiest way."

- "It would choose for the safe 2/3 marble option instead of the dangerous 0/1 or 4 marble option."

- "I made my choices based on how many marbles I could miss if the computer would turn left or right. In most cases I made the safe choice."

- "My thoughts were about which most profitable route the computer would take, by looking at how many marbles the computer would get in comparison to me. If they were even or less then I think the computer would play safely and take the best and safest option available at that point."

- "Look at the potential payoffs for blue in relation to the potential payoff for orange and check for probabilities."

6. REFERENCES

[1] I. Arieli and R. Aumann. The logic of backward induction. Available at SSRN: http://ssrn.com/abstract=2133302 or http://dx.doi.org/10.2139/ssrn.2133302, 2012.

[2] D. Balkenborg and R. Nagel. An experiment on forward versus backward induction: How fairness and levels of reasoning matter. Working paper, 2008.

[3] P. Battigalli. Strategic rationality orderings and the best rationalizability principle. *Games and Economic Behavior*, 13:178–200, 1996.

[4] P. Battigalli. On rationalizability in extensive games. *Journal of Economic Theory*, 74:40–61, 1997.

[5] G. Cachon and C. Camerer. Loss-avoidance and forward induction in experimental coordination games. *The Quarterly Journal of Economics*, 111(1):165–94, 1996.

[6] C. Camerer. *Behavioral Game Theory*. Princeton University Press, 2003.

[7] J. Chen and S. Micali. The robustness of extensive-form rationalizability. Working paper, 2011.

[8] J. Chen and S. Micali. The order independence of iterated dominance in extensive games. *Theoretical Economics*, 8:125–163, 2013.

[9] D. Fudenberg and E. Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica: Journal of the Econometric Society*, 54:533–554, 1986.

[10] S. Ghosh, B. Meijering, and R. Verbrugge. Strategic reasoning: Building cognitive models from logical formulas. *Journal of Logic, Language and Information*, 23(1):1–29, 2014.

[11] T. Halder, K. Sharma, S. Ghosh, and R. Verbrugge. How do adults reason about their opponent? Typologies of players in a turn-taking game. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 2015.

[12] T. Hedden and J. Zhang. What do you think i think you think?: Strategic reasoning in matrix games. *Cognition*, 85(1):1–36, 2002.

[13] A. Heifetz. *Game Theory: Interactive strategies in Economics and Management*. Cambridge University Press, 2012.

[14] A. Heifetz and A. Perea. On the outcome equivalence of backward induction and extensive form rationalizability. *International Journal of Game Theory*, 44:37–59, 2015.

[15] S. Huck and W. Müller. Burning money and (pseudo) first-mover advantages: An experimental study on forward induction. *Games and Economic Behavior*, 51(1):109–127, 2005.

[16] R. McKelvey and T. Palfrey. An experimental study of the centipede game. *Econometrica: Journal of the Econometric Society*, pages 803–836, 1992.

[17] B. Meijering, N. A. Taatgen, H. Van Rijn, and R. Verbrugge. Modeling inference of mental states: As simple as possible, as complex as necessary. *Interaction Studies*, 15:455–477, 2014.

[18] B. Meijering, H. Van Rijn, N. Taatgen, and R. Verbrugge. I do know what you think I think: Second-order theory of mind in strategic games is not that difficult. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 2486–2491, 2011.

[19] R. Nagel and F. Tang. Experimental results on the centipede game in normal form: An investigation on learning. *Journal of Mathematical Psychology*, 42(2):356–384, 1998.

[20] D. Pearce. Rationalizable strategic behaviour and the problem of perfection. *Econometrica*, 52:1029–1050, 1984.

[21] A. Perea. Epistemic foundations for backward induction: An overview. In J. van Benthem, D. Gabbay, and B. Löwe, editors, *Interactive Logic*, volume 1 of *Texts in Logic and Games*, pages 159–193. Amsterdam University Press, 2007.

[22] A. Perea. *Epistemic Game Theory: Reasoning and Choice*. Cambridge University Press, 2012.

[23] A. Perea. Belief in the opponents' future rationality. *Games and Economic Behavior*, 83:231–254, 2014.

[24] P. Reny. Backward induction, normal form perfection and explicable equilibria. *Econometrica*, 60:627–649, 1992.

[25] R. Rosenthal. Games of perfect information, predatory pricing, and the chain store. *Journal of Economic Theory*, 25(1):92–100, 1981.

[26] Q. Shahriar. An experimental test of the robustness and the power of forward induction. *Managerial and Decision Economics*, 35(4):264–277, 2014.

Ceteris paribus logic in counterfactual reasoning

Patrick Girard
The University of Auckland
Auckland, New Zealand

Marcus Anthony Triplett
The University of Auckland
Auckland, New Zealand

ABSTRACT

The semantics for counterfactuals due to David Lewis has been challenged on the basis of unlikely, or impossible, events. Such events may skew a given similarity order in favour of those possible worlds which exhibit them. By updating the relational structure of a model according to a *ceteris paribus* clause one forces out, in a natural manner, those possible worlds which do not satisfy the requirements of the clause. We develop a ceteris paribus logic for counterfactual reasoning capable of performing such actions, and offer several alternative (relaxed) interpretations of *ceteris paribus*. We apply this framework in a way which allows us to reason counterfactually without having our similarity order skewed by unlikely events. This continues the investigation of formal ceteris paribus reasoning, which has previously been applied to preferences [22], logics of game forms [9], and questions in decision-making [25], among other areas [16].

1. INTRODUCTION

The principal task of this paper is to work towards integrating *ceteris paribus* modalities into conditional logics so that some dissonant analyses of counterfactuals may be reconciled. We also suggest that ceteris paribus clauses may be understood dynamically, in the sense of dynamic epistemic logic [23], and we interpret our resulting ceteris paribus logic accordingly. Ceteris paribus clauses implicitly qualify many conditional statements that formulate laws of science and economics. A ceteris paribus clause adds to a statement a proviso requiring that other variables or states of affairs not explicitly mentioned in the statement are kept constant, thus ruling out benign defeaters. For instance, Avogadro's law says that if the volume of some ideal gas increases then, everything else held equal, the number of moles of that gas increases proportionally. Varying the temperature or pressure could provide situations that violate the plain statement of the law, but the ceteris paribus clause accounts for those. It specifically isolates the interaction between volume and number of moles by keeping everything else equal. In the same spirit, the Nash equilibrium in game theory is a solution concept that picks strategy profiles in which none of the agents could unilaterally (i.e., keeping the actions of others constant, or equal) deviate to their own advantage.

We may understand a ceteris paribus clause as a linguistic device intended to shrink the scope of the sentence qualified by the clause. For instance, when I make the utterance "I prefer fish to beef, *ceteris paribus*" I may mean something different from if I simply uttered "I prefer fish to beef." By enforcing the ceteris paribus condition I rule out some situations which affect my preference. For example if, whenever I eat fish I'm beaten with a mallet, while whenever I eat beef I'm left in peace, I might retract the second utterance and maintain the first. The ceteris paribus clause reduces the number of states of affairs under consideration. For modal logicians, 'ruling out' states of affairs amounts to strengthening an accessibility relation, consequently changing the relational structure of a model. This bears similarity to the epistemological forcing of Vincent Hendricks [10], which seeks to rule out 'irrelevant alternatives' in a way which allows knowledge in spite of the possibility of error. Wesley Holliday [11] develops several interpretations of the epistemic operator K based on the relevant alternatives epistemology; namely, that in order for an agent to have knowledge of a proposition, that agent must eliminate each *relevant alternative*. Holliday's semantics are based on the semantics for counterfactuals due to David Lewis [13], which we will recall in the next section. One could see relevant worlds as those which keep things equal. When reasoning using Avogadro's law, the relevant possible worlds are those where the temperature and pressure have not changed. Thus, in order for an agent to have knowledge, that agent must eliminate the alternatives among the worlds which 'keep things equal.'

Previously, ceteris paribus formalisms have been given for logics of preference [22] and logics of game forms [9]. Here we extend the analysis to counterfactual reasoning. The importance of counterfactuals in game theory is well known (see, for instance, [17]). For example, Bassel Tarbush [21] argues that the *Sure-Thing Principle*[1] ought to be understood as an inherently counterfactual notion. We will motivate our discussion by thinking through Kit Fine's well-known 'minor-miracles' argument [8], a putative counterexample to Lewis' semantics. We will argue that ceteris paribus logic, suitably adapted to conditionals, provides a natural response to this kind of argument. Moreover, we will see that ceteris paribus logic reveals a useful feature missing from the standard formalisation of counterfactuals; namely, the explicit requirement that certain propositions must have their truth remain fixed during the evaluation of the counterfactual. This is implicitly thought to hold, to some degree, when one works with models which have similarity orders or systems of spheres. The conditional logic of Graham Priest [15] makes just that assumption, but with no syntactic assurance. Ceteris paribus logic provides, in addition to the

[1] An outcome o of an action A is a *sure-thing* if, were any other action A' to be chosen, o would remain an outcome. The Sure-Thing Principle [18] states that sure-things should not affect an agent's preferences.

underlying similarity order over possible worlds, a syntactic apparatus to reason with such ceteris paribus clauses directly in the object language.

2. COUNTERFACTUALS

Here we shall formalise counterfactuals in the style of Lewis. Let Prop be a set of propositional variables. We are concerned with models of the form $\mathcal{M} = (W, \preceq, V)$ such that the following obtain.

1. W is a non-empty set of *possible worlds.*
2. $\preceq$ is a family $\{\preceq_w\}_{w \in W}$ of *similarity orders*, i.e., relations on $W_w \times W_w$ (with $W_w \subseteq W$) such that:
   - $w \in W_w$,
   - $\preceq_w$ is reflexive, transitive and total, and
   - $w \prec_w v$ for all $v \in W_w \setminus \{w\}$.
3. V is a *valuation function* assigning a subset $V(p) \subseteq W$ to each propositional variable $p \in \mathsf{Prop}$.

Intuitively, W_w is the set of worlds which are entertainable from w. Worlds which are not entertainable from w are deemed simply too dissimilar from w to be considered. Say that u is at least as similar to w as v is when $u \preceq_w v$, and that it is strictly more similar when $u \prec_w v$.

If $\mathcal{M}$ satisfies each of the three requirements we call $\mathcal{M}$ a *conditional model.* A relation $\leq$ is said to be *well-founded* if for every non-empty $S \subseteq W$ the set

$$\mathsf{Min}^{\mathcal{M}}_{\leq}(S) = \{v \in S \cap W : \text{ there is no } u \text{ with } u < v\} \quad (1)$$

is non-empty.[2] We will suppress the superscript $\mathcal{M}$ if it is clear from the context which model we're discussing. If a model $\mathcal{M} = (W, \preceq, V)$ has only well-founded similarity orders we say that $\mathcal{M}$ satisfies the *limit assumption.* For ease of exposition, we will assume that our conditional models satisfy the limit assumption. Of course, we may generalise the semantics for counterfactuals in the usual way [13], so that our results work for models which do not satisfy the limit assumption as well.

Definition 1 (Language $\mathcal{L}^{\Box\!\!\rightarrow}$). *The language $\mathcal{L}^{\Box\!\!\rightarrow}$ of counterfactuals is given by the following grammar*

$$\varphi \ ::= \ p \mid \neg\varphi \mid \varphi \vee \varphi \mid \varphi \mathrel{\Box\!\!\rightarrow} \psi.$$

We define $\varphi \wedge \psi := \neg(\neg\varphi \vee \neg\psi)$, $\varphi \rightarrow \psi := \neg\varphi \vee \psi$, $\varphi \mathrel{\Diamond\!\!\rightarrow} \psi := \neg(\varphi \mathrel{\Box\!\!\rightarrow} \neg\psi)$.

Definition 2 (Semantics). *Let $\mathcal{M} = (W, \preceq, V)$ be a well-founded conditional model. Then*

$$\begin{array}{rcl} [\![p]\!]^{\mathcal{M}} & = & V(p) \\ [\![\neg\varphi]\!]^{\mathcal{M}} & = & W \setminus [\![\varphi]\!]^{\mathcal{M}} \\ [\![\varphi \vee \psi]\!]^{\mathcal{M}} & = & [\![\varphi]\!]^{\mathcal{M}} \cup [\![\psi]\!]^{\mathcal{M}} \\ [\![\varphi \mathrel{\Box\!\!\rightarrow} \psi]\!]^{\mathcal{M}} & = & \{w \in W : \mathsf{Min}_{\preceq_w}([\![\varphi]\!]^{\mathcal{M}}) \subseteq [\![\psi]\!]^{\mathcal{M}}\}. \end{array}$$

Let $w \in W$. If $w \in [\![\varphi]\!]$ we write $\mathcal{M}, w \models \varphi$, and if $w \notin [\![\varphi]\!]$ we write $\mathcal{M}, w \not\models \varphi$.

3. THE NIXON ARGUMENT

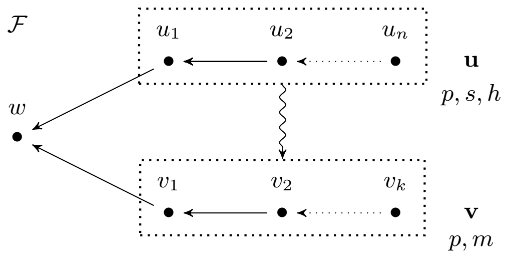
There is a problem dating back to the 1970s [1,3,8] surrounding the semantics for counterfactuals proposed by Lewis. We have found that our 'ceteris paribus counterfactuals' (defined below) provide a unique perspective on the problem (a putative counterexample). The argument goes as follows. Assume, during the Cold War, that President Richard Nixon had access to a device which launches a nuclear missile at the Soviets. All Nixon is required to do is press a button on the device. Consider the counterfactual *if Nixon had pushed the button, there would have been a nuclear holocaust.* Call it the *Nixon couterfactual.* It is not so difficult to see that the Nixon counterfactual could be true, or could be imagined to be true. Indeed, one could argue that the Nixon counterfactual ought to be true in any successful theory of counterfactuals. Fine and Lewis both agree (and so do we) that the counterfactual is true ([8, p. 452], [14, p. 468]), but Fine used the Nixon counterfactual to argue that the Lewis' semantics yields the wrong verdict. This is because "a world with a single miracle but no holocaust is closer to reality than one with a holocaust but no miracle." [8, p. 452] In response, Lewis argues that, provided the Nixon situation is modelled using a similarity relation which respects a plausible system of priorities (see below), the counterfactual will emerge true. We will provide a different response using ceteris paribus counterfactuals, but first let us see how Fine and Lewis model the situation.

Consider two classes of possible worlds. One class, $\mathbf{u}$, consists of those worlds in which Nixon pushes the button, and the button successfully launches the missile. The second, $\mathbf{v}$, consists of those worlds in which Nixon pushes the button, but some small occurrence – such as a minor miracle – prevents the button's correct operation. Certainly those worlds where the button does *not* launch the missile bear more similarity to the present world than those where it does. This is Fine's interpretation of Lewis' semantics. Any world in $\mathbf{u}$ has been devastated by nuclear warfare, countless lives have been lost, there is nuclear winter, etc., whereas worlds in $\mathbf{v}$ continue on as they would have done.

To illustrate Fine's interpretation, let p, s, m, h be the propositions:

p	=	"Nixon pushes the button,"
s	=	"the missile successfully launches,"
m	=	"a miracle prevents the missile being launched,"
h	=	"a nuclear holocaust occurs,"

and consider the following model, the *Fine model*:



An arrow from x to y indicates relative similarity to w, so u_1 is more similar to w than u_2 is. Arrows are transitive, and the 'snake' arrow between $\mathbf{u}$ and $\mathbf{v}$ indicates that $v_i \preceq_w u_j$ for every i, j. For each $u_i \in \mathbf{u}$, $\mathcal{F}, u_i \models p \wedge s \wedge h$; and for each $v_i \in \mathbf{v}$, $\mathcal{F}, v_i \models p \wedge m$. World w is intended to represent the

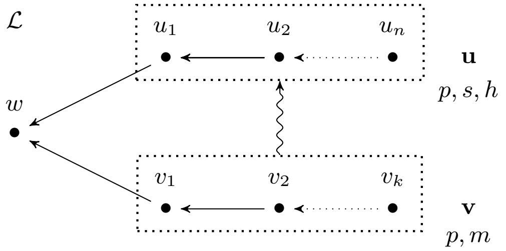
[2] As usual, $u < v$ is defined as $u \leq v$ and not $v \leq u$

real world: Nixon did not push any catastrophic anti-Soviet buttons,[3] no nuclear missile was successfully launched at the Soviets, no miracle prevented any such missile, and no nuclear holocaust occurred. World v_1 is more similar to w than any world in $\mathbf{u}$ is, since in any $\mathbf{u}$-world Nixon pushes the button and begins a nuclear holocaust. By (1), v_1 is therefore the minimal p-world. At v_1 the proposition h is false, and so $\mathcal{F}, w \not\models p \mathrel{\Box\!\!\rightarrow} h$. Therefore, Fine concludes, the Nixon counterfactual is false in Lewis' semantics.

In response, Lewis argues that the proper similarity relation to model the Nixon counterfactual should respect the following system of priorities:

1. It is of the first importance to avoid big, widespread, diverse violations of law.
2. It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
3. It is of the third importance to avoid even small, localized, simple violations of law.
4. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly. ([14, p. 472])

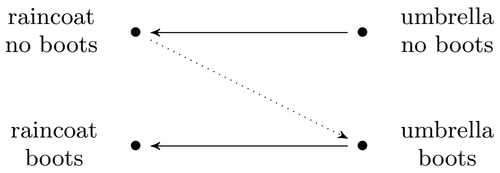
Based on this system of priorities world u_1 is more similar to w than v_1 is because "perfect match of particular fact counts for much more than imperfect match, even if the imperfect match is good enough to give us similarity in respects that matter very much to us." [14, p. 470] That is, worlds in $\mathbf{v}$ in which a small miracle prevents the missile being launched may look quite similar to our world, but only approximately so. And in Lewis' system of priorities, perfect match outweighs approximate similarity. The *Lewis model*, then, looks like this:



In the Lewis model, u_1 is the world most similar to w, and in u_1 the missile successfully launches, there is a nuclear holocaust, and so the Nixon counterfactual is true. Lewis thus responds to Fine by defending a similarity order that favours u_1 over v_1. He is justified by prioritising perfect over approximate match in a similarity relation according to the aforementioned system.

The interpretation of the Nixon counterfactual we will offer is in line with Lewis', though we do not rely on his system of priorities. We will achieve a resolution similar to his without having to defend a model different from Fine's. After all, as Lewis says: "I do not claim that this pre-eminence of perfect match is intuitively obvious. I do not claim that it is a feature of the similarity relations most likely to guide our explicit judgments. It is not; else the objection we are considering never would have been put forward."[14, p. 470] Instead, we will treat the Nixon counterfactual with an explicit ceteris paribus clause, dispatching with the unintuitive pre-eminence of perfect match in constructing the similarity relation.

Our interpretation of the Nixon counterfactual is much like in preference logic, where formal ceteris paribus reasoning was first applied [7, 22, 24]. Consider the following diagram, which shows a preference of a raincoat to an umbrella, provided wearing boots is kept constant:



Arrows point to more preferred alternatives, and are transitive. Evidently, having an umbrella and boots is preferred to having a raincoat and no boots. The variation of having boots *skews* the preference. If a ceteris paribus clause is enforced, guaranteeing that in either case boots will be worn or boots will not be worn, then the correct preference is recovered. A similar situation occurs in the logic of counterfactuals. The variation of certain propositions can skew the similarity order. In Fine's argument, this is done by the variation of physical law, a miracle. If we were to restrict the worlds considered during the evaluation of the counterfactual to those that agree with w on the proposition m, then in $\mathcal{F}$ the world v_1 would no longer assume the role of minimal p-world. Rather, u_1 would. In world u_1 a nuclear holocaust *does* occur, whence the counterfactual becomes true, as desired. This is our resolution of the Nixon argument, which we next formalise.

4. CETERIS PARIBUS SEMANTICS

We introduce into our language a new conditional operator which generalises the usual one. In particular, it accommodates explicit ceteris paribus clauses. The authors in [22] were the first to define object languages in this way. They developed a modal logic of ceteris paribus preferences in the sense of von Wright [24]. For now we will take the ordinary conditional operator and embed within it a finite set of formulas Γ understood as containing the *other things* to be kept equal.[4]

Definition 3 (Language $\mathcal{L}_{\mathsf{CP}}$). *Let Γ be a finite set of formulas. Then the language $\mathcal{L}_{\mathsf{CP}}$ is given by the grammar*[5]

$$\varphi \ ::= \ p \mid \neg\varphi \mid \varphi \vee \varphi \mid [\varphi, \Gamma]\psi.$$

We understand the modality $[\varphi, \Gamma]\psi$ as the counterfactual $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ subject to the requirement that the truth of the

[3] Although there is no way for us to know this, for the sake of the argument we assume that it is so.

[4] The choice of Γ finite is largely technical. We will mention some possibilities and difficulties regarding the case where the ceteris paribus set Γ may be infinite in our concluding remarks.

[5] We redefine the language more precisely as Definition 8 in the appendix. For simplicity we work with the one now stated.

formulas in Γ does not change. We define $\varphi \wedge \psi := \neg(\neg\varphi \vee \neg\psi)$, $\varphi \to \psi := \neg\varphi \vee \psi$, $\langle\varphi, \Gamma\rangle\psi := \neg[\varphi, \Gamma]\neg\psi$. We call the conditional $[\varphi, \Gamma]\psi$ a *ceteris paribus conditional*, or, if the antecedent is false, a *ceteris paribus counterfactual.* $\mathcal{L}_{\mathsf{CP}}$ is interpreted over standard conditional models, and thus requires no additional semantic information.

Some additional notation is required, however. Let $\mathcal{M} = (W, \preceq, V)$ be a conditional model and let $w, u, v \in \mathcal{M}$. Let $\Gamma \subseteq \mathcal{L}_{\mathsf{CP}}$ be finite.

- Define the relation $\equiv_\Gamma$ over W by $u \equiv_\Gamma v$ if for all $\gamma \in \Gamma$, $\mathcal{M}, u \models \gamma$ iff $\mathcal{M}, v \models \gamma$. Then $\equiv_\Gamma$ is an equivalence relation.[6]
- Set $[w]_\Gamma = \{u \in W_w : w \equiv_\Gamma u\}$, the collection of w-entertainable worlds which agree with w on Γ.
- Define $\trianglelefteq^\Gamma_w := \preceq_w \cap ([w]_\Gamma \times [w]_\Gamma)$, the restriction of $\preceq_w$ to the above worlds.

Thus if $u, v \in [w]_\Gamma$ then either $u \trianglelefteq^\Gamma_w v$ or $v \trianglelefteq^\Gamma_w u$.

Definition 4 (Semantics). *Let $\mathcal{M} = (W, \preceq, V)$ be a conditional model. Then*

$$[\![[\varphi, \Gamma]\psi]\!]^{\mathcal{M}} \quad = \quad \{w \in W : \mathsf{Min}_{\trianglelefteq^\Gamma_w}([\![\varphi]\!]^{\mathcal{M}}) \subseteq [\![\psi]\!]^{\mathcal{M}}\}.$$

The semantics for the regular connectives are the same as those in Definition 2. Notice that we recover the ordinary counterfactual $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ with $[\varphi, \emptyset]\psi$.

Consider again the Fine model $\mathcal{F}$. As before we have $\mathcal{F}, w \not\models p \mathrel{\Box\!\!\rightarrow} h$, but now

$$\mathcal{F}, w \models [p, \{m\}]h. \tag{2}$$

We thus think about the Nixon counterfactual by way of ceteris paribus reasoning. Allowing the truth of arbitrary formulas to vary during the evaluation of a counterfactual can distort the given similarity order, thereby attributing falsity to a sentence which may be intuitively true. By forcing certain formulas to keep their truth status fixed one can rule out these cases, which has just been demonstrated with (2). This ceteris paribus qualification is done in preference logic, and indeed in more general scientific and economic practice.[7] The Nixon counterfactual is simply a situation involving a defeater, or an irrelevant alternative, which ought to be forced out.

5. CETERIS PARIBUS AS A DYNAMIC ACTION

The modality $[\varphi, \Gamma]\psi$ behaves like a dynamic operator, in the sense of dynamic epistemic logic. For modality-free formulas φ and ψ, evaluating $[\varphi, \Gamma]\psi$ at $w \in W$ amounts to transforming

$$\mathcal{M} = (W, \{\preceq_w\}_{w \in W}, V)$$

into

$$[\Gamma]\mathcal{M} = (W, \{\trianglelefteq^\Gamma_w\}_{w \in W}, V)$$

and evaluating $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ at $[\Gamma]\mathcal{M}, w$. This dynamic action is possible since we are altering the relational structure of $\mathcal{M}$ with only a finite amount of information from Γ.

Note that the set W_w on which $\preceq_w$ is defined on may change after the update. By updating the model $\mathcal{M}$ with a ceteris paribus clause Γ, worlds which disagree on Γ are relegated to the class $W \setminus W_w$ of infinitely dissimilar (indeed, *irrelevant*) worlds. Figure 1 shows how the Fine model changes after being updated by a ceteris paribus clause forcing agreement on m. This forces out the $\mathbf{v}$-worlds from consideration during the evaluation of the counterfactual; in some sense syntactically 'correcting' the provided similarity order. Of course, if each world already agreed with w on $\{m\}$ the ceteris paribus clause would have no effect.

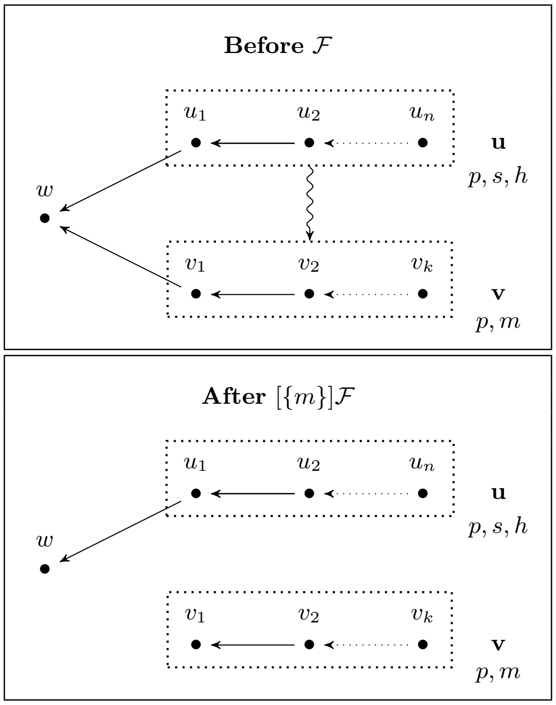


Figure 1: The Fine model before and after $\preceq_w$ is upgraded to $\trianglelefteq^{\{m\}}_w$.

The modality-free condition on φ and ψ cannot be removed. In particular, one cannot iterate the dynamic ceteris paribus action and retain agreement with the static ceteris paribus counterfactual operator. To see this, consider the example in Figure 2. Taking $\Gamma = \{s\}$ and $\Delta = \emptyset$, one has $\mathcal{M}, w \models [p, \Gamma][q, \Delta]r$, but $[\Gamma]\mathcal{M}, w \not\models p \mathrel{\Box\!\!\rightarrow} [q, \Delta]r$.

6. UNIFORMLY SELECTING CETERIS PARIBUS CLAUSES

Having created a formalism which accommodates explicit ceteris paribus clauses, one might desire a method for uniformly selecting the ceteris paribus set Γ. For von Wright [24], ceteris paribus means fixing every propositional variable which does not occur in the universe of discourse of the ceteris paribus expression under consideration. More precisely, let $\mathsf{UD}(\varphi)$ be the set of all propositional variables occurring in the formula φ, defined inductively as follows.

[6] Technically, the relation $\equiv_\Gamma$ should be defined together with the semantics in Definition 4 by mutual recursion. Again, we favour the simpler presentation.

[7] See Schurz [19] on comparative ceteris paribus laws.

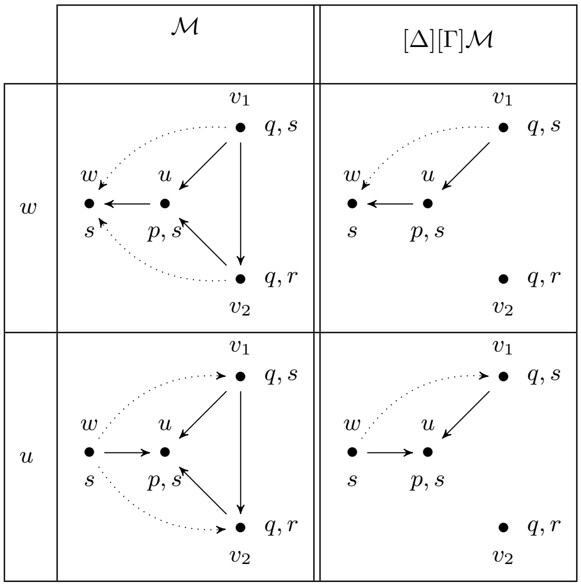


Figure 2: The horizontal panels labelled w and u define the similarity orders $\preceq_w$ and $\preceq_u$ respectively.

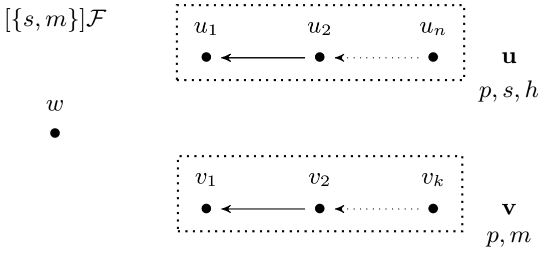
$$\begin{aligned}
\mathsf{UD}(p) &= \{p\}\\
\mathsf{UD}(\neg\varphi) &= \mathsf{UD}(\varphi)\\
\mathsf{UD}(\varphi \vee \psi) &= \mathsf{UD}(\varphi) \cup \mathsf{UD}(\psi)\\
\mathsf{UD}([\varphi, \Gamma]\psi) &= \mathsf{UD}(\varphi) \cup \mathsf{UD}(\Gamma) \cup \mathsf{UD}(\psi)\\
\mathsf{UD}(\{\gamma_1, \ldots, \gamma_n\}) &= \mathsf{UD}(\gamma_1) \cup \cdots \cup \mathsf{UD}(\gamma_n).
\end{aligned}$$

Then the ceteris paribus counterfactual *if φ were the case then, ceteris paribus, ψ would be the case* amounts to the expression

$$[\varphi, \mathsf{Prop} \setminus (\mathsf{UD}(\varphi) \cup \mathsf{UD}(\psi))]\psi. \qquad (3)$$

Now all propositional variables not occurring in the universe of discourse of the counterfactual antecedent or consequent are fixed.

Updating the Fine model with respect to von Wright's ceteris paribus set yields the following model:



We have $\mathcal{F}, w \models [p, \{m, s\}]h$, but vacuously! It appears that the relation $\trianglelefteq^\Gamma$ is too strong to interact with von Wright's definition. We are requiring that *everything else* is kept equal. This is questionable metaphysics, to say the least. Lewis made a similar observation in [13], about the counterfactual *'if kangaroos had no tails, they would topple over'*:

> We might think it best to confine our attention to worlds where kangaroos have no tails and *everything* else is as it actually is; but there are no such worlds. Are we to suppose that kangaroos have no tails but that their tracks in the sand are as they actually are? Then we shall have to suppose that these tracks are produced in a way quite different from the actual way. [...] Are we to suppose that kangaroos have no tails but that their genetic makeup is as it actually is? Then we shall have to suppose that genes control growth in a way quite different from the actual way (or else that there is something, unlike anything there actually is, that removes the tails). And so it goes; respects of similarity and difference trade off. If we try too hard for exact similarity to the actual world in one respect, we will get excessive differences in some other respect. ([13, p. 9])

In fact, for the logic of *ceteris paribus* counterfactuals to function in a meaningful fashion, every formula occurring in Γ must be independent from the counterfactual antecedent. In the Fine model, we insist that the truth values of s and m are kept fixed. These propositions, however, are nomologically related to p, so we can't change the truth value of p without affecting the truth values of s and m. This is why the counterfactual $[p, \{m, s\}]h$ is vacuously true, but then so is the counterfactual $[p, \{m, s\}]\neg h$. To accommodate a uniform method for selecting ceteris paribus clauses, more flexibility is required. What *ought* to be kept equal when we can't keep *everything else* equal? In the next section we will consider two strategies for relaxing the interpretation of *ceteris paribus* to address this question.

7. RELAXING THE CETERIS PARIBUS CLAUSE

7.1 Naïve counting

We will now introduce another interpretation for the modality $[\varphi, \Gamma]\psi$. Let us write $[\![[\varphi, \Gamma]\psi]\!]^{\mathcal{M}}_{\mathsf{CP}}$ for the set $[\![[\varphi, \Gamma]\psi]\!]^{\mathcal{M}}$ from Definition 4, and let $\models_{\mathsf{CP}}$ act as the ordinary satisfaction relation for Boolean formulas, but with

$$\mathcal{M}, w \models_{\mathsf{CP}} [\varphi, \Gamma]\psi \text{ iff } w \in [\![[\varphi, \Gamma]\psi]\!]^{\mathcal{M}}_{\mathsf{CP}}.$$

Whereas in Definition 4 we required strict agreement on the set Γ, in order to develop a logic for ceteris paribus counterfactuals with a weaker semantics we will instead relax the requirement to *maximal agreement*. The best we can do is preserve the set Γ as much as possible for any given model.

Let $\Gamma \subseteq \mathcal{L}_{\mathsf{CP}}$ be finite, and let $\mathcal{M} = (W, \preceq, V)$ be a conditional model. Define $A^{\mathcal{M}}_\Gamma : W \times W \to 2^\Gamma$ by

$$A^{\mathcal{M}}_\Gamma(u, v) = \{\gamma \in \Gamma : \mathcal{M}, u \models \gamma \text{ iff } \mathcal{M}, v \models \gamma\}. \qquad (4)$$

Define the relation $\preceq^\Gamma_w$ on W_w by $u \preceq^\Gamma_w v$ iff

$$\begin{aligned}
&\text{either } |A^{\mathcal{M}}_\Gamma(u, w)| > |A^{\mathcal{M}}_\Gamma(v, w)|,\\
&\quad\text{or } |A^{\mathcal{M}}_\Gamma(u, w)| = |A^{\mathcal{M}}_\Gamma(v, w)| \text{ and } u \preceq_w v.
\end{aligned}$$

The relation $\preceq^\Gamma_w$ can be seen as a transformed $\preceq_w$, reordering the similarity order so that worlds closer to w preserve at least as much of Γ as worlds further away, and if any two worlds agree on Γ to the same quantity, then the nearer world is more similar to w with respect to $\preceq$.

Definition 5 (Semantics). *Let $\mathcal{M} = (W, \preceq, V)$ be a conditional model satisfying the limit assumption. Let $\Gamma \subseteq$*

$\mathcal{L}_{\mathsf{CP}}$ *be finite. Then*

$$[\![[\varphi, \Gamma]\psi]\!]^{\mathcal{M}}_{\mathsf{NC}} \quad = \quad \{w \in W : \mathsf{Min}_{\preceq^{\Gamma}_{w}}([\![\varphi]\!]^{\mathcal{M}}) \subseteq [\![\psi]\!]^{\mathcal{M}}\}.$$

We write $\mathcal{M}, w \models_{\mathsf{NC}} [\varphi, \Gamma]\psi$ iff $w \in [\![[\varphi, \Gamma]\psi]\!]^{\mathcal{M}}_{\mathsf{NC}}$.

FACT 1. *Let $\mathcal{M} = (W, \preceq, V)$ be a conditional model. Let $w \in W$, and let $\mathbf{X} \in \{\mathsf{CP}, \mathsf{NC}\}$. Then the following are true, where $\pm\alpha$ is shorthand which uniformly stands for either α or $\neg\alpha$:*

1. $\mathcal{M}, w \models \varphi \Box\!\!\rightarrow \psi$ *iff* $\mathcal{M}, w \models_{\mathbf{X}} [\varphi, \emptyset]\psi$
2. $\mathcal{M}, w \models_{\mathbf{X}} (\pm\alpha \wedge \langle\varphi, \Gamma\rangle(\pm\alpha \wedge \psi)) \rightarrow \langle\varphi, \Gamma \cup \{\alpha\}\rangle\psi$
3. $\mathcal{M}, w \models_{\mathsf{CP}} \langle\varphi, \Gamma\rangle\psi \Rightarrow \mathcal{M}, w \models_{\mathsf{NC}} \langle\varphi, \Gamma\rangle\psi$
4. $\mathcal{M}, w \models_{\mathsf{NC}} [\varphi, \Gamma]\psi \Rightarrow \mathcal{M}, w \models_{\mathsf{CP}} [\varphi, \Gamma]\psi$

The original ceteris paribus preference logic [22] could be axiomatised using standard axioms together with Fact 1.2 and its converse. A crucial difference with NC semantics is that the converse of Fact 1.2 does not hold. The existence of a $\varphi \wedge \psi$-world which maximally agrees on $\Gamma \cup \{\alpha\}$ does not ensure that α actually holds at that world. In fact, it is not guaranteed that *any* formula from $\Gamma \cup \{\alpha\}$ is obtained.

7.2 Maximal supersets

An approach to counterfactuals familiar to the AI community [4–6,12] makes use of a selection function which chooses the 'closest' world according to maximal sets of propositional variables. More specifically, each world w satisfies some set $\mathbf{P}_w \subseteq \mathsf{Prop}$ of propositional variables, and a world u is a world closest to w if there is no v with $\mathbf{P}_u \subset \mathbf{P}_v \subseteq \mathbf{P}_w$. Taking this as a kind of ceteris paribus formalism we obtain the following variant of our ceteris paribus counterfactuals. First let us define the relation $\sqsubseteq^{\Gamma}_{w}$ on W_w by $u \sqsubseteq^{\Gamma}_{w} v$ iff

$$\begin{aligned} &\text{either } A^{\mathcal{M}}_{\Gamma}(v, w) \subset A^{\mathcal{M}}_{\Gamma}(u, w), \\ &\quad\text{or } A^{\mathcal{M}}_{\Gamma}(v, w) = A^{\mathcal{M}}_{\Gamma}(u, w) \text{ and } u \preceq_w v. \end{aligned}$$

DEFINITION 6 (SEMANTICS). *Let $\mathcal{M} = (W, \preceq, V)$ be a conditional model satisfying the limit assumption. Let $\Gamma \subset \mathcal{L}_{\mathsf{CP}}$ be finite. Then*

$$[\![[\varphi, \Gamma]\psi]\!]^{\mathcal{M}}_{\mathsf{MS}} \quad = \quad \{w \in W : \mathsf{Min}_{\sqsubseteq^{\Gamma}_{w}}([\![\varphi]\!]^{\mathcal{M}}) \subseteq [\![\psi]\!]^{\mathcal{M}}\}.$$

We write $\mathcal{M}, w \models_{\mathsf{MS}} [\varphi, \Gamma]\psi$ iff $w \in [\![[\varphi, \Gamma]\psi]\!]^{\mathcal{M}}_{\mathsf{MS}}$. Now Γ is maximally preserved in the sense that worlds which preserve the same propositions as another, and furthermore preserve additional propositions from Γ, are deemed to approximate Γ more closely; while worlds u, v with neither $A^{\mathcal{M}}_{\Gamma}(u, w) \subseteq A^{\mathcal{M}}_{\Gamma}(v, w)$ nor $A^{\mathcal{M}}_{\Gamma}(v, w) \subseteq A^{\mathcal{M}}_{\Gamma}(u, w)$ are considered incomparable.

FACT 2 (EXTENDS FACT 1). *Let $\mathcal{M} = (W, \preceq, V)$ be a conditional model. Let $w \in W$. Then the following are true.*

1. $\mathcal{M}, w \models \varphi \Box\!\!\rightarrow \psi$ *iff* $\mathcal{M}, w \models_{\mathsf{MS}} [\varphi, \emptyset]\psi$
2. $\mathcal{M}, w \models_{\mathsf{MS}} (\pm\alpha \wedge \langle\varphi, \Gamma\rangle(\pm\alpha \wedge \psi)) \rightarrow \langle\varphi, \Gamma \cup \{\alpha\}\rangle\psi$
3. $\mathcal{M}, w \models_{\mathsf{CP}} \langle\varphi, \Gamma\rangle\psi \Rightarrow \mathcal{M}, w \models_{\mathsf{MS}} \langle\varphi, \Gamma\rangle\psi$
4. $\mathcal{M}, w \models_{\mathsf{MS}} [\varphi, \Gamma]\psi \Rightarrow \mathcal{M}, w \models_{\mathsf{CP}} [\varphi, \Gamma]\psi$

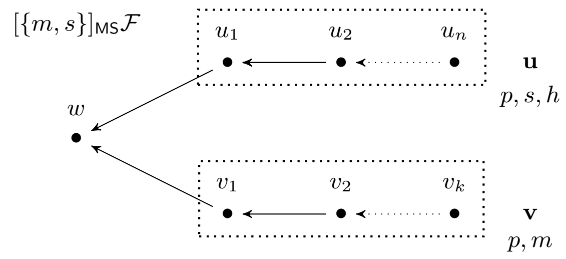
8. DYNAMICS AND THE NIXON COUNTERFACTUAL

Given a ceteris paribus interpretation $\mathbf{X} \in \{\mathsf{CP}, \mathsf{NC}, \mathsf{MS}\}$, let us write $[\Gamma]_{\mathbf{X}}\mathcal{M}$ for the model $\mathcal{M}$ updated with a ceteris paribus clause Γ according to interpretation $\mathbf{X}$. Specifically, we have the following definition.

DEFINITION 7. *Let $\mathcal{M} = (W, \preceq, V)$ be a conditional model, and let $\Gamma \subseteq \mathcal{L}_{\mathsf{CP}}$ be a finite set of formulas. We define the* updated models *$[\Gamma]_{\mathbf{X}}\mathcal{M}$, for $\mathbf{X} \in \{\mathsf{CP}, \mathsf{NC}, \mathsf{MS}\}$, by*

$$\begin{aligned} [\Gamma]_{\mathsf{CP}}\mathcal{M} &:= (W, \trianglelefteq^{\Gamma}, V); \\ [\Gamma]_{\mathsf{NC}}\mathcal{M} &:= (W, \preceq^{\Gamma}, V); \\ [\Gamma]_{\mathsf{MS}}\mathcal{M} &:= (W, \sqsubseteq^{\Gamma}, V). \end{aligned}$$

This provides us with three dynamic *ceteris paribus* updates. Let us see how they treat the Nixon counterfactual. We have already witnessed the CP update with ceteris paribus sets $\{m\}$ and $\{m, s\}$, and concluded that both make the counterfactual true (vacuous truth with $\{m, s\}$). NC and MS updates agree on the truth of the Nixon counterfactual with the CP update on $\{m\}$, but disagree on $\{m, s\}$. Updating the Fine model with von Wright's ceteris paribus clause $\{m, s\}$ according to the NC interpretation yields $\mathcal{F}$ again. Thus $\mathcal{F}, w \not\models_{\mathsf{NC}} [p, \{m, s\}]h$. Updating Fine's model with $\{m, s\}$ according to the MS interpretation gives the following model:



In $[\{m, s\}]_{\mathsf{MS}}\mathcal{F}$ the Nixon counterfactual is not true, and neither is $p \Box\!\!\rightarrow \neg h$.

We summarise the truth of the Nixon counterfactuals $p \Box\!\!\rightarrow h$ and $p \Box\!\!\rightarrow \neg h$ in the various updated Fine models in the following table.

		Interpretation		
Counterfactual	*Clause*	CP	NC	MS
$p \Box\!\!\rightarrow h$	$\{m\}$	true	true	true
	$\{m, s\}$	true	false	false
$p \Box\!\!\rightarrow \neg h$	$\{m\}$	false	false	false
	$\{m, s\}$	true	true	false

The rows labelled with $p \Box\!\!\rightarrow h$ and $p \Box\!\!\rightarrow \neg h$ indicate the truth value of those counterfactuals in the updated models $[\Gamma]_{\mathbf{X}}\mathcal{F}$, where Γ is given by the cell in the *Clause* column and $\mathbf{X}$ is given by the *Interpretation* column.

Formally, the table illustrates how different truth values for the Nixon counterfactual may be obtained by combining the various interpretations of ceteris paribus (CP, NC, MS) with the different ceteris paribus sets (the selected set $\{m\}$ or von Wright's set $\{m, s\}$). But this doesn't mean that all combinations are legitimate formalisations of Fine's argument. Fine's story is about small miracles that can interfere with Nixon's ploy, not about whether the missile would successfully launch should Nixon press the button. That the proposition s must be able to vary is crucial to the story, so one shouldn't attempt to keep it equal, on a par with m. We adhere to our favoured formalisation of the Nixon argument in which the proposition m is the only one that needs to be kept equal. We have given principled reasons for this choice, and our selection makes the counterfactual true – all interpretations agree on that. The point of the table is a formal one, namely that the truth-values of counterfactuals vary with different ceteris paribus updates according to their interpretation.

9. THEOREMS

In the appendix (Corollary 1) we prove that the logic $\Lambda^{\mathcal{L}_{\mathsf{CP}}}_{\mathfrak{C}}$ of ceteris paribus counterfactuals over the class of conditional frames $\mathfrak{C}$ is complete for $\mathsf{CP}/\mathsf{NC}/\mathsf{MS}$ semantics. The proof works by translating formulas of $\mathcal{L}_{\mathsf{CP}}$ into formulas of a *comparative possibility* language, in the style of Lewis, and axiomatising the equivalent logic. This permits a clearer reduction of ceteris paribus modalities to basic comparative possibility operators, albeit with a translation exponential in the size of Γ.

10. CONCLUDING REMARKS

This paper has introduced a ceteris paribus logic for counterfactual reasoning by adapting the formalism in [22]. We have introduced some variants on ceteris paribus logic in light of philosophical difficulties arising in the application of conditionals. We apply our framework to the *Nixon counterfactual*, and with this bring a new perspective to the problem. We have suggested and explored the dynamic perspective of our various syntactic interpretations of *ceteris paribus*, which has resulted in a richer understanding of so-called *comparative* ceteris paribus reasoning in formal settings. We have provided completeness theorems which demonstrate that the ceteris paribus logics so obtained ultimately reduce to the underlying counterfactual logic; in our case Lewis' VC. With our framework we defend Lewisian semantics by appealing to examples from preference logic, where ceteris paribus reasoning is more widely discussed.

Finally, we outline some limitations of our framework and directions for future research.

Iterated ceteris paribus actions. We saw in Section 5 that iterated ceteris paribus counterfactuals deviate in truth from the corresponding update-then-counterfactual sequence. While this is undesirable, it is not so uncommon to face such technical difficulties with iterated counterfactuals. It remains to further understand the interaction between the two.

Cardinality restrictions on Γ. In general, ceteris paribus reasoning requires keeping equal as much information as possible, and sometimes unknown information (for example, unanticipated defeaters of laws). Keeping everything else equal may indeed mean keeping equal an indefinite, and possibly infinite, set of things. Exploring ceteris paribus logic without cardinality restrictions to Γ is thus more than a mere technical exercise. But it is not so straightforward to extend the present framework to accommodate the presence of infinite Γ. The translations presented in the appendix only carry over to the infinite case for infinitary languages, which is not much of a solution. For the strict ceteris paribus semantics, we instead suggest following the δ-flexibility approach of [20]. For the relaxed ceteris paribus semantics, there are conceptual difficulties which arise with the comparison of infinite sets: when should we say of two infinite sets that one keeps more things equal than the other? Clearly naïve counting will not suffice. Minimising distance with respect to $\sqsubseteq^{\Gamma}$ is more promising, but has it's own problems. We leave this challenging technical enterprise for future research.

11. ACKNOWLEDGMENTS

We wish to thank the participants at the Australasian Association of Logic and the Analysis, Randomness and Applications meetings held in New Zealand in 2014, and in particular Sam Baron for valuable feedback on a previous version of this paper. We also wish to thank anonymous referees for valuable comments.

12. REFERENCES

[1] Jonathan Bennett. Review of Lewis ([13]). *The Canadian Journal of Philosophy*, 4:381–402, 1974.

[2] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Cambridge University Press, New York, NY, USA, 2001.

[3] G. Lee Bowie. The similarity approach to counterfactuals: Some problems. *Noûs*, pages 477–498, 1979.

[4] Mukesh Dalal. Investigations into a theory of knowledge base revision: preliminary report. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, volume 2, pages 475–479, 1988.

[5] Luis Fariñas del Cerro and Andreas Herzig. Interference logic = conditional logic + frame axiom. *International Journal of Intelligent Systems*, 9(1):119–130, 1994.

[6] Luis Fariñas del Cerro and Andreas Herzig. Belief change and dependence. In *Proceedings of the 6th conference on Theoretical aspects of rationality and knowledge*, pages 147–161. Morgan Kaufmann Publishers Inc., 1996.

[7] Jon Doyle and Michael P. Wellman. Representing preferences as ceteris paribus comparatives. *Ann Arbor*, 1001:48109–2110, 1994.

[8] Kit Fine. Review of Lewis' counterfactuals. *Mind*, 84:451–458, 1975.

[9] Davide Grossi, Emiliano Lorini, and Francois Schwarzentruber. Ceteris paribus structure in logics of game forms. *Proceedings of the 14th conference on theoretical aspects of rationality and knowledge. ACM.*, 2013.

[10] Vincent F. Hendricks. *Mainstream and formal epistemology*. Cambridge University Press, 2006.

[11] Wesley H. Holliday. Epistemic closure and epistemic logic i: Relevant alternatives and subjunctivism. *Journal of Philosophical Logic*, pages 1–62, 2014.

[12] Hirofumi Katsuno and Alberto O. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52(3):263–294, 1991.

[13] David Lewis. *Counterfactuals*. Harvard University Press, 1973.

[14] David Lewis. Counterfactual dependence and time's arrow. *Noûs*, pages 455–476, 1979.

[15] Graham Priest. *An introduction to non-classical logic: From if to is*. Cambridge University Press, 2008.

[16] Carlo Proietti and Gabriel Sandu. Fitch's paradox and ceteris paribus modalities. *Synthese*, 173(1):75–87, 2010.

[17] Dov Samet. Hypothetical knowledge and games with perfect information. *Games and economic behavior*, 17(2):230–251, 1996.

[18] Leonard J. Savage. *The foundations of statistics*. Courier Corporation, 1972.

[19] Gerhard Schurz. Ceteris paribus laws: Classification and deconstruction. *Erkenntnis (1975-)*, 57(3):pp. 351–372, 2002.

[20] Jeremy Seligman and Patrick Girard. Flexibility in ceteris paribus reasoning. *The Australasian Journal of Logic*, 10(0), 2011.

[21] Bassel Tarbush. Agreeing on decisions: an analysis with counterfactuals. *Proceedings of the 14th conference on theoretical aspects of rationality and knowledge. ACM.*, 2013.

[22] Johan van Benthem, Patrick Girard, and Olivier Roy. Everything else being equal: A modal logic for ceteris paribus preferences. *Journal of Philosophical Logic*, 38(1):83–125, 2009.

[23] Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Pieter Kooi. *Dynamic epistemic logic*, volume 337. Springer, 2007.

[24] Georg H. von Wright. *The Logic of Preference*. Edinburgh University Press, 1963.

[25] Zuojun Xiong and Jeremy Seligman. Open and closed questions in decision-making. *Electronic Notes in Theoretical Computer Science*, 278:261–274, 2011.

APPENDIX

We first recast Definition 3 in a more formally precise manner.

Definition 8. *For each ordinal α let $\mathcal{L}_\alpha$ be given by*

$$\varphi \ ::= \ p \mid \bot \mid \neg\varphi \mid \varphi \vee \psi \mid [\varphi, \Gamma]\psi$$

where $\Gamma \subseteq \mathcal{L}_\beta$ is finite and $\beta < \alpha$. $\mathcal{L}_{\mathsf{CP}}$ is then defined to be $\bigcup_\alpha \mathcal{L}_\alpha$.

This ensures the sets Γ are well-defined. One can define a language $\mathcal{L}$ of comparative possibility in a similar style, though we will only give the following grammar

$$\begin{aligned} \varphi \ ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \varphi \preceq \psi \mid \varphi \preceq^\Gamma \psi \mid \varphi \trianglelefteq^\Gamma \psi \\ \mid \varphi \sqsubseteq^\Gamma \psi. \end{aligned}$$

We further set

$$\begin{array}{lll lll} \varphi \prec \psi & := & \neg(\psi \preceq \varphi); & \varphi \prec^\Gamma \psi & := & \neg(\psi \preceq^\Gamma \varphi); \\ \varphi \lhd^\Gamma \psi & := & \neg(\psi \trianglelefteq^\Gamma \varphi); & \varphi \sqsubset^\Gamma \psi & := & \neg(\psi \sqsubseteq^\Gamma \varphi); \\ \Diamond\varphi & := & \varphi \prec \bot; & \Box\varphi & := & \neg\Diamond\neg\varphi. \end{array}$$

Definition 9 (Semantics). *Let $\mathcal{M}, w$ be a conditional model. Then*

$$\begin{array}{rcl} \llbracket p \rrbracket^{\mathcal{M}} & = & V(p); \\ \llbracket \neg\varphi \rrbracket^{\mathcal{M}} & = & W \setminus \llbracket \varphi \rrbracket^{\mathcal{M}}; \\ \llbracket \varphi \vee \psi \rrbracket^{\mathcal{M}} & = & \llbracket \varphi \rrbracket^{\mathcal{M}} \cup \llbracket \psi \rrbracket^{\mathcal{M}}; \\ \llbracket \varphi \preceq \psi \rrbracket^{\mathcal{M}} & = & \{w \in W : \forall u \in W_w\ \exists v \in W_w \text{ such that} \\ & & \text{if } u \in \llbracket \psi \rrbracket^{\mathcal{M}} \text{ then } v \in \llbracket \varphi \rrbracket^{\mathcal{M}} \text{ and } v \preceq_w u\}; \\ \llbracket \varphi \preceq^\Gamma \psi \rrbracket^{\mathcal{M}} & = & \{w \in W : \forall u \in W_w\ \exists v \in W_w \text{ such that} \\ & & \text{if } u \in \llbracket \psi \rrbracket^{\mathcal{M}} \text{ then } v \in \llbracket \varphi \rrbracket^{\mathcal{M}} \text{ and } v \preceq^\Gamma_w u\}; \\ \llbracket \varphi \trianglelefteq^\Gamma \psi \rrbracket^{\mathcal{M}} & = & \{w \in W : \forall u \in [w]_\Gamma\ \exists v \in [w]_\Gamma \text{ such that} \\ & & \text{if } u \in \llbracket \psi \rrbracket^{\mathcal{M}} \text{ then } v \in \llbracket \varphi \rrbracket^{\mathcal{M}} \text{ and } v \trianglelefteq^\Gamma_w u\}; \\ \llbracket \varphi \sqsubseteq^\Gamma \psi \rrbracket^{\mathcal{M}} & = & \{w \in W : \forall u \in W_w\ \exists v \in W_w \text{ such that} \\ & & \text{if } u \in \llbracket \psi \rrbracket^{\mathcal{M}} \text{ then } v \in \llbracket \varphi \rrbracket^{\mathcal{M}} \text{ and } v \sqsubseteq^\Gamma_w u\}. \end{array}$$

Lemma 1. *The modal operator $[\varphi, \Gamma]\psi$ under* NC *semantics is definable in $\mathcal{L}$.*

Proof. We show that

$$\mathfrak{M}, w \models_{\mathsf{NC}} [\varphi, \Gamma]\psi \text{ iff } \mathfrak{M}, w \models \Diamond\varphi \to (\varphi \wedge \psi) \prec^\Gamma (\varphi \wedge \neg\psi).$$

$\Rightarrow$: Assume $\mathcal{M}, w \models \Diamond\varphi$. Then there is a world $x \in W_w$ such that $\mathcal{M}, x \models \varphi$. By assumption there exists $y \in W_w$ such that $y \preceq^\Gamma_w x$, $\mathcal{M}, y \models \varphi$ and for all $z \preceq^\Gamma_w y$, we have $\mathcal{M}, z \models \varphi \to \psi$. Since $y \preceq^\Gamma_w y$, we obtain $\mathcal{M}, y \models \varphi \wedge \psi$. So there is $u \in W_w$ such that for all $v \in W_w$, if $\mathcal{M}, u \models \varphi \wedge \psi$ and $v \preceq^\Gamma_w u$ then $\mathcal{M}, v \not\models \varphi \wedge \neg\psi$. This is exactly $\mathcal{M}, w \models (\varphi \wedge \psi) \prec^\Gamma (\varphi \wedge \neg\psi)$.

$\Leftarrow$: By contrapositive. Assume $\mathcal{M}, w \not\models [\varphi, \Gamma]\psi$. By the semantic definition there exists $x \in W_w$ such that $\mathcal{M}, x \models \varphi$ and for all $y \in W_w$ with $y \preceq^\Gamma_w x$ and $\mathcal{M}, y \models \varphi$, there is $z \in W_w$ such that $z \preceq^\Gamma_w y$ and $\mathcal{M}, z \models \varphi \wedge \neg\psi$. Since $\mathcal{M}, x \models \varphi$ we have $\mathcal{M}, w \models \Diamond\varphi$. Now, take an arbitrary world $u \in W_w$ such that $\mathcal{M}, u \models \varphi \wedge \psi$. Either (i) $u \preceq^\Gamma_w x$ or (ii) $x \preceq^\Gamma_w u$. If (i), then the fact that $\mathcal{M}, u \models \varphi$ and $u \preceq^\Gamma_w x$ implies the existence of $u' \in W_w$ such that $u' \preceq^\Gamma_w u \preceq^\Gamma_w x$ and $\mathcal{M}, u' \models \varphi \wedge \neg\psi$. If (ii), then $\mathcal{M}, x \models \varphi$ and $x \preceq^\Gamma_w x$ together imply that there exists $x' \in W_w$ with $x' \preceq^\Gamma_w x \preceq^\Gamma_w u$ and $\mathcal{M}, x' \models \varphi \wedge \neg\psi$. Either way we have that for all $u \in W_w$ there exists $v \in W_w$ such that if $\mathcal{M}, u \models \varphi \wedge \psi$ then $v \preceq^\Gamma_w u$ and $\mathcal{M}, v \models \varphi \wedge \neg\psi$, whence $\mathcal{M}, w \not\models (\varphi \wedge \psi) \prec^\Gamma (\varphi \wedge \neg\psi)$. □

Lemma 2. *The modal operator $[\varphi, \Gamma]\psi$ under* CP *semantics is definable in $\mathcal{L}$.*

Proof. Replace $\preceq^\Gamma_w$ with $\trianglelefteq^\Gamma_w$ in the above proof to show that the following equivalence

$$\mathfrak{M}, w \models_{\mathsf{CP}} [\varphi, \Gamma]\psi \text{ iff } \mathfrak{M}, w \models \Diamond\varphi \to (\varphi \wedge \psi) \lhd^\Gamma (\varphi \wedge \neg\psi).$$

holds. □

Lemma 3. *The modal operator $[\varphi, \Gamma]\psi$ under* MS *semantics is definable in $\mathcal{L}$.*

Proof. Replace $\preceq^\Gamma_w$ with $\sqsubseteq^\Gamma_w$ in the above proof to show that the following equivalence holds

$$\mathfrak{M}, w \models_{\mathsf{MS}} [\varphi, \Gamma]\psi \text{ iff } \mathfrak{M}, w \models \Diamond\varphi \to (\varphi \wedge \psi) \sqsubset^\Gamma (\varphi \wedge \neg\psi).$$

□

Denote by $\mathcal{L}^-$ the $\mathcal{L}$-fragment given by

$$\varphi \ ::= \ p \in \mathsf{Prop} \mid \neg\varphi \mid \varphi \vee \varphi \mid \varphi \preceq \psi.$$

Given a set $\Gamma \subseteq \mathcal{L}$ or $\Gamma \subseteq \mathcal{L}^-$, let $\mathbf{\Gamma}$ be the set of all possible conjunctions of formulas and negated formulas from Γ; that is, the set of all ψ such that $\psi = \bigwedge_{\gamma \in \Gamma} \pm\gamma$, where $+\gamma = \gamma$ and $-\gamma = \neg\gamma$. So if $\Gamma = \{p, \neg q\}$ then

$$\mathbf{\Gamma} = \{p \wedge q, \neg p \wedge q, p \wedge \neg q, \neg p \wedge \neg q\}.$$

We will often identity a conjunction $\varphi_1 \wedge \cdots \wedge \varphi_n$ with the set $\{\varphi_1, \ldots, \varphi_n\}$.

Lemma 4. *The modal operator $\trianglelefteq^\Gamma$ of $\mathcal{L}$ is expressible in $\mathcal{L}^-$.*

Proof. We show that

$$\varphi \trianglelefteq^\Gamma \psi \leftrightarrow \bigwedge_{\overline{\Gamma} \in \mathbf{\Gamma}} \left[\overline{\Gamma} \to (\varphi \wedge \overline{\Gamma}) \preceq (\psi \wedge \overline{\Gamma})\right]. \tag{5}$$

$\Rightarrow$: Without loss of generality write $\mathcal{M}, w \models \overline{\Gamma}$. Let $u \in W_w$ and suppose $\mathcal{M}, u \models \psi \wedge \overline{\Gamma}$. By hypothesis there exists $v \in [w]_\Gamma$ such that $\mathcal{M}, v \models \varphi$ and $v \trianglelefteq^\Gamma_w u$. Now $v \equiv_\Gamma w$, so $\mathcal{M}, v \models \overline{\Gamma}$, and $v \preceq_w u$ as required.

$\Leftarrow$: Write $\mathcal{M}, w \models \overline{\Gamma}$. Then $\mathcal{M}, w \models (\varphi \wedge \overline{\Gamma}) \preceq (\psi \wedge \overline{\Gamma})$. Let $u \in [w]_\Gamma$ and suppose that $\mathcal{M}, u \models \psi$. Then $\mathcal{M}, u \models \psi \wedge \overline{\Gamma}$, so there exists $v \in W_w$ with $\mathcal{M}, v \models \varphi \wedge \overline{\Gamma}$ and $v \preceq_w u$. Then $v \equiv_\Gamma w$, and so $v \trianglelefteq^\Gamma_w u$. □

Lemma 5. *The modal operator $\sqsubseteq^\Gamma$ of $\mathcal{L}$ is expressible in $\mathcal{L}^-$.*

Proof. We show that

$$\varphi \sqsubseteq^\Gamma \psi \leftrightarrow \bigwedge_{\overline{\Gamma} \in \mathbf{\Gamma}} (\overline{\Gamma} \to \bigwedge_{\Lambda \subseteq \overline{\Gamma}} ([(\varphi \wedge \Lambda) \vee (\bigvee_{\Lambda \subset \Sigma \subseteq \overline{\Gamma}} \Sigma \wedge \varphi)] \preceq \psi \wedge \Lambda)). \tag{6}$$

$\Rightarrow$: Suppose $\mathcal{M}, w \models \varphi \sqsubseteq^\Gamma \psi$ with $\mathcal{M}, w \models \overline{\Gamma}$, for some $\overline{\Gamma} \in \mathbf{\Gamma}$. Let $u \in W_w$ be arbitrary and $\Lambda \subseteq \overline{\Gamma}$ such that $\mathcal{M}, u \models \psi \wedge \Lambda$. Then by hypothesis there is $v \in W_w$ such that $v \sqsubseteq^\Gamma_w u$ and $\mathcal{M}, v \models \varphi$. We have the following two cases.

Case 1: $A^{\mathcal{M}}_\Gamma(w,u) \subset A^{\mathcal{M}}_\Gamma(w,v)$.

Then since $\Lambda \subseteq A^{\mathcal{M}}_\Gamma(w,u)$ and $\mathcal{M}, v \models A^{\mathcal{M}}_\Gamma(w,v) \wedge \varphi$ we have

$$\mathcal{M}, v \models \bigvee_{\Lambda \subset \Sigma \subseteq \overline{\Gamma}} \Sigma \wedge \varphi$$

which shows the implication.

Case 2: $A^{\mathcal{M}}_\Gamma(w,u) = A^{\mathcal{M}}_\Gamma(w,v)$ and $v \preceq_w u$.

Then $\mathcal{M}, v \models \Lambda \wedge \varphi$, which shows the implication.

$\Leftarrow$: Let $u \in W_w$ be arbitrary such that $\mathcal{M}, u \models \psi$. Then $\mathcal{M}, u \models A^{\mathcal{M}}_\Gamma(w,u) \wedge \psi$. Let $\overline{\Gamma}$ be the unique element of $\mathbf{\Gamma}$ such that $\mathcal{M}, u \models \overline{\Gamma}$. By hypothesis there exists $v \in W_w$ such that $v \preceq_w u$ and

$$\mathcal{M}, v \models (\varphi \wedge A^{\mathcal{M}}_\Gamma(w,u)) \vee (\bigvee_{A^{\mathcal{M}}_\Gamma(w,u) \subset \Sigma \subseteq \overline{\Gamma}} \Sigma \wedge \varphi). \tag{7}$$

If the second disjunct from (7) holds then there is Σ with $A^{\mathcal{M}}_\Gamma(w,u) \subset \Sigma \subseteq \overline{\Gamma}$ such that $\mathcal{M}, v \models \Sigma \wedge \varphi$. Hence $A^{\mathcal{M}}_\Gamma(w,u) \subset A^{\mathcal{M}}_\Gamma(w,v)$. Thus $u \sqsubseteq^\Gamma_w v$, as required.

If the second disjunct from (7) fails then $\mathcal{M}, v \models \varphi \wedge A^{\mathcal{M}}_\Gamma(w,u)$. In particular $A^{\mathcal{M}}_\Gamma(w,u) = A^{\mathcal{M}}_\Gamma(w,v)$, otherwise the second disjunct would be true. Combining this with $v \preceq_w u$ we have $v \sqsubseteq^\Gamma_w u$, as required. □

Lemma 6. *The modal operator $\preceq^\Gamma$ of $\mathcal{L}$ is expressible in $\mathcal{L}^-$.*

Proof. Replace the subset condition

$$\bigvee_{\Lambda \subset \Sigma \subseteq \overline{\Gamma}} \Sigma \wedge \varphi$$

in (6) with the cardinality condition

$$\bigvee_{|\Lambda| < |\Sigma| \leq |\overline{\Gamma}|} \Sigma \wedge \varphi$$

and repeat the above process. □

Notice that, if $\Gamma \cup \{\varphi, \psi\} \subseteq \mathcal{L}^-$, then the right hand sides of the equivalences established above are in $\mathcal{L}^-$. This allows us to apply the translation to a formula from the inside-out, the resulting formula belonging to $\mathcal{L}^-$.

By a *conditional frame* we mean a pair $F = (W, \preceq)$, such that (F, V) is a conditional model for any valuation function V. Let $\mathfrak{C}$ be the class of conditional frames. Using the notation from [2], we write $\Lambda^{\mathfrak{L}}_{\mathfrak{C}}$ for the set of $\mathfrak{L}$-formulas valid over $\mathfrak{C}$.

Theorem 1. *The logic $\Lambda^{\mathcal{L}}_{\mathfrak{C}}$ is complete.*

Proof. We take as our axiomatisation the axioms for VC [13], plus the translations from Lemmas 4, 5, and 6. □

Corollary 1. *The logic $\Lambda^{\mathcal{L}_{\mathsf{CP}}}_{\mathfrak{C}}$ is complete for* CP/NC/MS-*semantics.*

An Information-Gap Framework for Capturing Preferences About Uncertainty

Russell Golman
Department of Social and Decision Sciences
Carnegie Mellon University
rgolman@andrew.cmu.edu

George Loewenstein
Department of Social and Decision Sciences
Carnegie Mellon University
gl20@andrew.cmu.edu

ABSTRACT

We propose an integrated theoretical framework that captures preferences for acquiring or avoiding information as well as preferences for exposure to uncertainty (i.e., risk or ambiguity) by allowing utility to depend not just on material payoffs but also on beliefs and the attention devoted to them. We use this framework to introduce the concept of an information gap – a specific uncertainty that one recognizes and is aware of. We characterize a specific utility function that describes feelings about information gaps. We suggest that feelings about information gaps are the source of curiosity as well as a second motive to manage one's thoughts through information acquisition or avoidance. In addition, we suggest that feelings about information gaps also contribute to risk- and ambiguity preferences.

Keywords

Ambiguity, curiosity, information gap, motivated attention, ostrich effect, risk

1. INTRODUCTION

In a seminal paper titled "The Mind as a Consuming Organ," Thomas Schelling (1987) pointed out that much consumption is not of the material sort, but takes place largely "in the mind." Research in psychology, decision theory, and economics has identified a number of motives underlying informational consumption, from the powerful force of curiosity (Loewenstein, 1994) to the pleasures of knowledge and insight (Karlsson et al., 2004). Moreover, even when missing information is not available to an individual, demand for this information plays a role in decision making under uncertainty. Here we propose a unified theoretical framework that allows us to model feelings about information and about *information gaps* – specific uncertainties that an individual recognizes and is aware of. We present a specific utility function that takes as input beliefs and the attention devoted to them (as well as material payoffs). This utility model can be applied to decision making about information acquisition or avoidance as well as to decision making under risk and ambiguity (as described in Table 1).

In one branch of the economics literature, preferences about information have been viewed as derivative from risk preferences (e.g., Kreps and Porteus, 1978; Wakker, 1988; Grant et al., 1998; Dillenberger, 2010). We take a complementary perspective, considering preferences about information as primitive and viewing preferences about risk and ambiguity as derivative of them.

Decision about:	Domain of:
Whether to address an uncertainty	Information acquisition or avoidance
Whether to expose oneself to an uncertainty	Risky or ambiguous choice

Table 1: Two domains of decision making affected by feelings about uncertainty.

The standard account of preferences about information holds that information is valuable because, and only to the extent that, it enables people to make superior decisions that raise their expected utility (Hirshleifer and Riley, 1979). Often, however, individuals seek information purely to satisfy curiosity, which refers to the desire for information for its own sake – i.e., specifically *not* for its ability to improve decision making. Curiosity correlates with brain activity in regions thought to relate to anticipated reward (Kang et al., 2009), suggesting that information is a reward in and of itself. Loewenstein (1994) proposed an information-gap account of curiosity, and our framework allows us to capture this motive for information acquisition within an expanded utility model. While curiosity is a powerful motive for information acquisition, there nevertheless are many situations in which people actively choose to avoid information, e.g., not obtaining a costless medical test. We hypothesize that information avoidance derives from a desire to avoid increasing attention on a negative anticipated outcome. More generally, we suggest that individuals have an inclination to seek (or avoid) information whenever they anticipate that what they discover will be pleasurable (or painful). Of course, ex-ante beliefs about such events are already good or bad respectively (Eliaz and Spiegler, 2006), but there can be a big difference between discovering something for sure and simply considering it a likely possibility. Our additional assumption is that obtaining news tends to increase attention to it (as in Gabaix et al., 2006; Tasoff and Madarász, 2009), which leads to the implication that people will seek information about questions they like thinking about and will avoid information about questions they do not like thinking about. We explore the implications of our proposed utility model for information acquisition or avoidance in a companion paper (Golman and Loewenstein, 2015a), and we outline this analysis in Section 5.

The standard account of preferences about risk and ambiguity considers these preferences to be primitives in a model (e.g., Anscombe and Aumann, 1963; Klibanoff et al., 2005). However, research has shown that missing information has a profound impact on decision making under risk and ambiguity. For example, Ritov and Baron (1990) studied hypothetical decisions concerning whether to vaccinate a child, when the vaccine reduces the risk of the child dying from a disease but might itself be harmful. When the uncertainty was caused by salient missing information about the risks from vaccination – a child had a high risk of being harmed by the vaccine or no risk at all but it was impossible to find out which – subjects were more reluctant to vaccinate than in a situation in which all children faced a similar risk and there was no salient missing information. In a second companion paper (Golman and Loewenstein, 2015b) we argue that the information-gap concept developed here underlies an alternative account of risk and ambiguity aversion (and seeking) that is conceptually different from, and has different testable implications from, the usual account of risk aversion involving loss aversion and the usual account of ambiguity aversion involving vague probabilities.[1] In Section 6 we outline our argument that salient information gaps can either increase or decrease preference for uncertain gambles depending on whether it is painful or pleasurable to think about the information one is missing.

Our expanded utility model builds on the insights of Caplin and Leahy (2001).[2] Caplin and Leahy recognize that anticipatory feelings about prizes that might be received in the future can affect utility. We follow them (and Kőszegi (2010) as well) in applying expected utility theory to psychological states rather than to physical prizes, but we expand the domain of psychological states that people can have feelings about. In doing so, we incorporate Tasoff and Madarász's (2009) insight that information stimulates attention and thus complements anticipatory feelings. Kreps and Porteus (1978) present a model capturing preferences for early or late resolution of uncertainty, and Dillenberger (2010) captures preferences for one-shot or sequential resolution of uncertainty; this line of research thus deals with when, but not whether, an individual prefers to acquire information. Our model focuses just on the latter issue, but with it one could address the timing of uncertainty resolution by making additional assumptions about time preference.

We rely on a reduced form model of knowledge and awareness to describe information gaps – and the desire to fill them or ignore them – in order to avoid the complications of working with information partitions in a state-space model of knowledge (as in Aumann, 1976). The standard partitional state-space framework permits a distinction between two states of affairs – knowing and not knowing – but makes it difficult to capture unawareness (Modica and Rustichini, 1994; Dekel et al., 1998). We introduce a question-answer knowledge structure that allows us easily to draw an important distinction between *three* different states: knowing (represented by a question and a particular answer); not knowing, but knowing that one doesn't know (represented by a question and a set of possible answers); and not knowing and not knowing what one doesn't know (represented by the absence of an activated question). This third state corresponds to pure unawareness (Li, 2008), in the sense that an individual is unaware of the question itself and does not even distinguish different possible answers. (In contrast, our question-answer structure does not capture partial unawareness, in the sense of an individual being aware of a question and proper subset of possible answers, but unaware of some other remaining possible answers.) The question-answer structure is consistent with, and could be cast in terms of, a generalized state-space model (e.g., Modica and Rustichini, 1999; Heifetz et al., 2006), but we find the question-answer structure convenient to use.

The question-answer knowledge structure is intended to reflect human information-processing capabilities. Our cognitive maps of the world are not sets of possible states, each described in exquisite detail to account for all possible consequences of all possible decisions. Instead, people attend to a few relevant aspects of a situation and use limited information to make a broad judgment that can be refined later, if necessary. People tend to set goals and monitor their progress toward them in order to navigate a complex world (Miller et al., 1960; Locke and Latham, 1990; Loewenstein, 1999). We advance the idea that the acquisition of knowledge is also goal-oriented. We don't simply seek out information to maximize the data available to us or even to optimize future decisions, but instead tend to seek answers to questions that are either posed to us or that we pose to ourselves. Questions are, therefore, very much like informational goals or reference points. Indeed, focusing on a question that one cannot answer – e.g., a puzzle one cannot figure out – can torment a person and at the same time motivate the search for an answer, much as a high reference point can simultaneously detract from utility and motivate one to strive to reach it.

2. THEORETICAL FRAMEWORK

2.1 Cognitive States

Traditional economic theory assumes that utility is a function of consumption bundles or material outcomes, or (perhaps subjective) distributions thereof. Our basic premise is that utility depends not only on such material outcomes but also on one's cognitive state, encompassing the attention paid to each of the issues or questions that one is aware of as well as subjective judgments about the possible answers to these questions. While people have preferences about their beliefs (and the attention paid to them), we do not treat beliefs (or attention) as choice variables. People can choose whether or not to acquire information that will influence beliefs, but we assume that one's beliefs, given one's information, are constrained by Bayesian inference.

While there surely is an infinite set of possible states of the world, we assume, realistically we believe, that a person can only conceive of a finite number of questions at any one time. We represent awareness with an array of '*activated*' questions and a remaining set of '*latent*' questions. Activated questions are those that the individual is aware of. Latent questions are those that the individual could become, but is not currently, aware of. The finite subset of questions a person is aware of (i.e., paying at least some attention to) is denoted $\mathcal{Q}$. We label these activated questions as $Q_1, \ldots, Q_m$. A vector of attention weights $\mathbf{w} = (w_1, \ldots, w_m) \in \mathbb{R}^m_+$ indicates how much attention each activated question gets.[3] These

[1] For example, we show that low-stakes risk aversion (Rabin, 2000) could be attributed to the discomfort of thinking about uncertainties.

[2] Many have considered the notion that people derive utility from their beliefs (Abelson, 1986; Geanakoplos et al., 1989; Asch et al., 1990; Yariv, 2001; Kadane et al., 2008).

[3] We can think of the (presumably infinite) set of latent ques-

attention weights depend on three factors that we designate "*importance*," "*salience*," and "*surprise*." We return to define and discuss these concepts in Section 3.

A question Q_i has a countable set[4] of possible (mutually exclusive) answers $\mathcal{A}_i = \{A_i^1, A_i^2, \ldots\}$.[5] A person may not know the correct answer to a given question, but reasonably has a subjective belief about the probability that each answer is correct.[6] (The subjective probabilities across different questions may well be mutually dependent.) This framework allows us to capture information gaps, which are represented as activated questions lacking known correct answers, as depicted in Table 2.

Question	Answer	Belief	
Latent	–	Unawareness	
Activated	Unknown	Uncertainty	↕ information gap
	Known	Certainty	

Table 2: The question-answer knowledge structure.

Anticipated material outcomes, or prizes, can also be incorporated into this framework. We let X denote a countable set of prizes – i.e., material outcomes. The subjective probability over these prizes is in general mutually dependent with the subjective probability over answers to activated questions; that is, the receipt of new information often leads to revised beliefs about the likelihood of answers to many different questions as well as about the likelihood of different material outcomes. Denote the space of answer sets together with prizes as $\alpha = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_m \times X$. Thus, given a state of awareness defined by the set of activated questions $\mathcal{Q}$,[7] we represent a person's cognitive state C with a subjective probability measure π defined over α (i.e., over possible answers to activated questions as well as eventual prizes) and a vector of attention weights $\mathbf{w}$. We denote the set of all possible cognitive states as $\mathcal{C} = \Delta(\alpha) \times \mathbb{R}_+^m$ (with the notation $\Delta(\alpha)$ referring to the space of probability distributions over α with finite entropy. The restriction to distributions with finite entropy serves a technical purpose, but it should not trouble us – intuitively, it means that a person cannot be aware of an infinite amount of information, which is also the basis for our assumption that the set of activated questions is finite.). Each marginal distribution π_i specifies the subjective probability of possible answers to question Q_i, and similarly π_X specifies the subjective probability over prizes.[8]

The formal representation of a cognitive state is depicted in Table 3. Consider, for example, a college professor deciding whether or not to look at her teaching ratings. The set of activated questions (and possible answers) might include: "How many of my students liked my teaching?" $(0, 1, 2, \ldots)$; "Did they applaud on the last day of class?" (yes/no); "How good a teacher am I?" (great, good, so-so, bad, awful); "Will I get tenure?" (yes/no). Prior belief about the first question might be quite uncertain. The answer to the second question, on the other hand, might already be known with certainty. There may or may not be much uncertainty about the third and fourth questions. All of these beliefs (to the extent they are uncertain) are jointly dependent. The material outcome might be next year's salary, which would also depend on (but not be completely determined by) whether or not she gets tenure. Looking at the ratings will definitively answer the first question and may resolve some, but not all, of the uncertainty surrounding the other issues.

2.2 Actions

A decision maker has the possibility of taking actions with two kinds of effects: *informational* actions contribute to subjective judgments about the world by answering a question; and *instrumental* actions affect the chances of receiving various prizes (outcomes). For example, wagering on the color of a ball drawn from an urn is an instrumental action. Examining the contents of the urn is an informational action. Informational actions affect the subjective probability measure through the conditioning of beliefs on the discovered answer. Instrumental actions affect beliefs directly by changing the distribution over prizes conditional on subjective judgments. Both instrumental and informational actions also impact attention weights through their respective effects on importance and surprise. Note that some actions will have both instrumental and informational effects. Examples include paying a fee for a property value appraisal or hiring a private eye.

At any point in time an individual can be characterized by a prior cognitive state consisting of subjective probability measure π^0 and attention weight vector $\mathbf{w}^0$. Actions, in general, are operators on cognitive states that map to new cognitive states or to distributions over cognitive states. A purely instrumental action acting on the prior cognitive state determines a particular new cognitive state. Typically, it preserves the prior subjective judgment about the probability of each answer set and then specifies a new distribution over prizes conditional on each possible answer set. An instrumental action may also affect the importance of various questions (as formalized in the next section) and thereby influence the attention weights. For example, the decision to participate in a karaoke session will likely raise the attention weight on the question "Am I a good singer?"

Acquiring information also changes one's cognitive state. Ex ante, as one does not know which answer will be discovered, the prospect of acquiring information offers the decision maker a lottery over cognitive states. Upon learning answer A_i to question Q_i, one's subjective probability measure over $\Delta(\alpha)$ changes from π^0 to $\pi^{A_i} = \pi^0(\cdot|A_i)$.[9] We assume Bayesian updating here, which means that ex ante, before one knows what one will discover, an informational action determines a distribution over subjective judgments such that the expectation of this distribution equals the prior judgment. That is, by the law of total probability,

tions as having attention weights of zero.

[4] We use the term countable here to mean *at most countable*. The restriction of a countable set of answers to a countable set of possible questions does still allow an uncountable set of possible states of the world, but as awareness is finite, the precise state of the world would be unknowable.

[5] We assume that there is no such thing as an answer that is disconnected from a question.

[6] By *subjective* probability, we mean personal probability, but we take it to be observable by direct elicitation.

[7] In most cases, we will assume that activation of questions is determined exogenously – i.e., by the environment. We don't model growing awareness (see Karni and Vierø, 2013).

[8] For any $\tilde{\mathcal{A}} \subseteq \mathcal{A}_i$, we have $\pi_i(\tilde{\mathcal{A}}) = \pi(\mathcal{A}_1 \times \cdots \times \mathcal{A}_{i-1} \times \tilde{\mathcal{A}} \times \mathcal{A}_{i+1} \times \cdots \times \mathcal{A}_m \times X)$.

[9] We thus denote a belief with complete certainty in $\mathbf{A}\times x$ as $\pi^{\mathbf{A}\times x}$.

Activated Questions	Possible Answers	Subjective Probabilities[*]	Attention Weights
Q_1	$\mathcal{A}_1 = \{A_1^1, A_1^2, \ldots\}$	$[\pi_1(A_1^1), \pi_1(A_1^2), \ldots]$	w_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Q_m	$\mathcal{A}_m = \{A_m^1, A_m^2, \ldots\}$	$[\pi_m(A_m^1), \pi_m(A_m^2), \ldots]$	w_m
	Possible Prizes		
N/A	$X = \{x, x', x'', \ldots\}$	$[\pi_X(x), \pi_X(x'), \ldots]$	N/A

[*] Answers to different questions are not generally independent. Typically, the joint probability measure $\pi \neq \pi_1 \cdots \pi_m \cdot \pi_X$.

Table 3: Representation of a cognitive state.

$\sum_{A_i \in \mathcal{A}_i} \pi_i^0(A_i)\, \pi^{A_i} = \pi^0$. An informational action would decrease expected entropy because conditioning reduces entropy (see, e.g., Cover and Thomas, 1991, pg. 27). New information generates surprise (as formalized in the next section), which changes the attention weights too. Given the prior attention weight vector $\mathbf{w}^0$ based on salience and importance, we let $\mathbf{w}^{A_i}$ denote the new attention weight vector immediately after learning A_i, resulting from surprise at this discovery.

2.3 Preferences over (Distributions of) Cognitive States

The conventional theory of choice under risk assumes that a lottery over outcomes is evaluated according to its expected utility. Given that we may think of an informational action as creating a lottery over cognitive states, we make the natural assumptions leading to an expected utility representation in this new domain.

Independence Across Cognitive States

We assume that *there is a complete and transitive preference relation $\succeq$ on $\Delta(\mathcal{C})$ that is continuous (with respect to an appropriate topology)*[10] *and that satisfies independence*, so there exists a continuous expected utility representation u of $\succeq$ (von Neumann and Morgenstern, 1944).

The assumption here is that when information could put a person into one of many possible cognitive states, preference is consistent with valuing each possible cognitive state independently of any other cognitive states the person might have found herself in.

This might seem to imply that the utility of a state of uncertain knowledge is equal to the expected utility of each of the possible beliefs – e.g., that being uncertain of whether the object of my desire reciprocates my affections provides the same utility as the sum of probabilities times the utilities associated with the possible outcome belief states. It need not, because (as we discuss in detail below) obtaining the information, and indeed the specific information one obtains, is likely to affect one's attention weights. Such a change in attention can encourage or discourage a decision maker from resolving uncertainty, depending on whether the news that will be revealed is expected to be good or bad.

2.4 Choosing Between Sequences of Actions

The discovery of information following an initial action can change the availability or desirability of subsequent actions. For example, the information in a college professor's teaching ratings could help her decide whether to enroll in a teacher improvement class. A sequence of actions can be analyzed with the convention that an action operator passes through a distribution over cognitive states.[11] Thus, we represent a sequence of actions s acting on a cognitive state $(\pi, \mathbf{w})$ as $s\cdot(\pi, \mathbf{w}) \in \Delta(\mathcal{C})$.

Choice from among a set of sequences of actions $\mathcal{S}$, where early actions may reveal information that will inform later actions, is represented as utility maximization: a sequence $s^* \in \mathcal{S}$ may be chosen by a decision maker in the cognitive state $(\pi, \mathbf{w})$ if $s^* \in \arg\max_{s \in \mathcal{S}} u\,(s\cdot(\pi, \mathbf{w}))$. We find it useful to define a utility function over cognitive states, contingent on the set of sequences of actions that may subsequently be chosen:

$$U(\pi, \mathbf{w} \,|\, \mathcal{S}) = \max_{s \in \mathcal{S}} u\,(s\cdot(\pi, \mathbf{w}))\,. \qquad (1)$$

In the example of the professor's teaching ratings, the set of available subsequent actions is to enroll in the teacher improvement class or not to enroll in the class. Looking at the ratings resolves a lottery over cognitive states, each of which having utility that is conditional on making the optimal choice of one of these subsequent actions.

We define the desirability of a sequence of actions s in cognitive state $(\pi, \mathbf{w})$ as $D(s \,|\, \pi, \mathbf{w}) = u\,(s\cdot(\pi, \mathbf{w})) - u\,(\pi, \mathbf{w})$.[12] Desirability is simply marginal utility relative to the trivial 'action' of doing nothing.

3. PSYCHOLOGICAL INSIGHTS

In this section we introduce a number of specific psychological insights that lead us to specify a utility function that generates a wide range of testable predictions concerning informational phenomena. These insights help us characterize the factors that influence the level of attention paid to a question as well as to identify distinctly the valence of beliefs and the desire for clarity.

3.1 Attention

Neuroeconomic research indicates that attention shapes preference (Fehr and Rangel, 2011). Attention weights in our model specify how much a person is thinking about particular beliefs and, in turn, how much those beliefs directly impact utility. We may think of beliefs as having intrinsic value, which is then amplified by these attention weights.

[10] The induced topology on $\mathcal{C}$ (derived from the order topology on $\Delta(\mathcal{C})$) should be a refinement of the order topology on $\mathcal{C}$ (see Nielsen, 1984).

[11] Analogous to the standard assumption in decision under risk, the model assumes reduction of compound distributions over cognitive states. This does not imply the traditional reduction of compound lotteries.

[12] The degenerate distributions in $\Delta(\mathcal{C})$ correspond to individual states of knowledge. With the standard abuse of notation, we refer to the utility of the degenerate distribution on $(\pi, \mathbf{w}) \in \mathcal{C}$ as $u(\pi, \mathbf{w})$.

Our model (assuming monotonicity with respect to attention weights, as described in Section 4) provides a natural distinction between beliefs that have positive or negative intrinsic value: beliefs are positive specifically when more attention enhances utility and are negative in the opposite case. That is, a person likes thinking about (i.e., putting more attention weight on) *positive beliefs* and does not like thinking about *negative beliefs.*

Here we formalize the concepts of *importance*, *salience*, and *surprise*, all of which, we assume, contribute to attention weight. The importance γ_i of a question Q_i reflects the degree to which one's utility depends on the answer. Thus, for example, for an egocentric, but insecure, individual, the question, "Do other people like me?" is likely to be of great importance because the answer matters to the individual. Salience, distinctly, reflects the degree to which a particular context highlights the question. If, for example, an individual hears that another person was talking about her (with no further details), the question of whether the comments were favorable or not will become highly salient. We denote the salience of question Q_i as $\sigma_i \in \mathbb{R}_+$. Finally, surprise is a factor that reflects the dependence of attention on the dynamics of information revelation, and specifically on the degree to which receiving new information changes one's beliefs. If, having believed that she was generally well-liked, our individual were to discover that the comments about her were actually unfavorable, the discovery, necessitating a radical change in her belief, would be quite surprising (and, as we presently assume, would increase her attention to the question). We denote the surprise associated with a revised belief about question Q_i as δ_i. We assume that *the attention w_i on an activated question Q_i is a strictly increasing function of this question's importance γ_i, its salience σ_i, and the surprise δ_i associated with it.*

Importance

The importance of a question depends on the spread of the utilities associated with the different answers to that question. The degree to which an individual's utility varies with the answers to a question depends both on the magnitude of the utility function and on the perceived likelihood of different answers. Continuing with the example of the question of how well-liked an individual is, one could distinguish two relevant traits: egocentrism – the degree to which the individual *cares* about being well-liked; and insecurity – the dispersion of the individual's subjective probability distribution across possible answers. By our definition of the concept, importance should be positively related to both traits.

Given a particular prior subjective probability measure π^0 and a set $\mathcal{S}$ of sequences of actions available to the decision maker, the importance γ_i of question Q_i is a function (only) of the likelihood of possible answers and the utilities associated with these answers, captured as

$$\gamma_i = \phi\left(\left\langle \pi_i^0(A_i),\, U\big(\pi^{A_i}, \mathbf{w}^{A_i} \,|\, \mathcal{S}\big)\right\rangle_{A_i \in \operatorname{supp}\left(\pi_i^0\right)}\right)$$

where U is the utility function defined in Equation (1). Without specifying the precise form of this function ϕ, we assume only that *it (i.e., importance) increases with mean-preserving spreads of the (subjective) distribution of utilities that would result from different answers to the question*, and that *it is invariant with respect to constant shifts of utility.* Thus, a question is important to the extent that one's utility depends on the answer. Raising the stakes increases importance. On the other hand, if an answer is known with certainty, then by this definition nothing is at stake, so the underlying question is no longer important. While acquiring information will affect the importance of the questions being addressed, it takes time to adapt to news, so there should be some delay. We assume that *the importance of a question is updated only when the new information is incorporated into a new default subjective probability measure.*

Our definition of importance is, admittedly, circular. Importance depends on utility, which in turn depends on the attention weight, but importance also contributes to attention weight. There is, likely, some psychological realism to this circularity which captures the dynamic processes giving rise to obsession: attention to a question raises its importance, and the elevated importance gives rise to intensified attention. If we assume that these processes unfold instantaneously, then importance (and, in turn, attention weight and utility) will be a fixed point of this composition of functions. We can make simple comparisons of importance without going to the trouble of specifying precise values.

Salience

The salience of a question depends on a variety of exogenous contextual factors. For example, a question could be salient if it has recently come up in conversation (i.e., it has been primed) or if other aspects of the environment remind an individual about it. Alternatively, a question could be more salient to an individual if the answer is, in principle, knowable, and even more so if other people around her know the answer but she does not.

Often a question may be salient despite being unimportant. Continuing the prior example, even if an individual deems others' perceptions of her as unimportant, the question of her popularity might nonetheless be highly salient if the individual was asked, "Do you know what x thinks of you?" Conversely, there are myriad questions that are important by the definition just provided, but which lack salience. There might be numerous people whose opinion of us we would care about and be unsure of, but unless something raises the issue in our mind, we are unlikely to focus on it. It seems natural to think that some degree of salience is a necessary, and sufficient, condition for attention (while some degree of importance is not). Thus, we assume that a question Q_i is activated (i.e., has strictly positive attention weight $w_i > 0$) if and only if it has positive salience $\sigma_i > 0$. Further, we assume that *attention weight w_i has strictly increasing differences (i.e., a positive cross-partial derivative, if we assume differentiability) in* (γ_i, σ_i). That is, an increase in importance produces a greater increase in attention for a more salient question.

Surprise

The third factor that we posit influences attention is the surprise one experiences upon acquiring new information. Surprise reflects the degree to which new information changes existing beliefs. A natural measure of surprise was proposed in a theoretical paper by Baldi (2002) and, in an empirical follow-up investigation (Itti and Baldi, 2009), shown to predict the level of attention paid to information. Incorporating the insights from this line of research, we assume that *when the answer to a particular question Q_j is learned, thereby contributing information about the answers to asso-*

ciated questions and causing their subjective probabilities to be updated, the degree of surprise associated with a new belief about question Q_i can be defined as the Kullback-Leibler divergence of $\pi_i^{A_j}$ against the prior π_i^0:

$$\delta_i(\pi_i^{A_j} || \pi_i^0) = \sum_{A_i \in \mathcal{A}_i} \pi_i^{A_j}(A_i) \log \frac{\pi_i^{A_j}(A_i)}{\pi_i^0(A_i)}.$$

Surprise is positive with any new information, and is greatest when one learns the most unexpected answer with certainty. However, the feeling of surprise is not permanent. We assume that *when the decision maker adapts and gets used to this new knowledge (formally, when the default subjective probability measure is reset), it is no longer surprising.*

The Belief Resolution Effect

The impact of new information on attention is greatest when uncertainty about a question is resolved completely. Surprise immediately spikes, but in the long run fades, and the underlying question becomes unimportant because, with the answer known, there is no longer a range of possible answers. Taken together, these factors create a pattern of change in attention weight following the discovery of a definitive answer, what we call the *belief resolution effect* – when an answer is learned with certainty, there is an immediate boost in attention weight on it, but over time this attention weight falls to a lower level. Specifically, when the decision maker adapts and the certain belief is incorporated into the default subjective probability measure, the question then receives less attention. It is as if the brain recognizes that because a question has been answered, it can move on to other questions that have yet to be addressed. Janis (1958) recognized the belief resolution effect when he observed that surgical patients getting information about their upcoming procedures initially worry more about the surgery but subsequently experience less anxiety.

3.2 Valence and Clarity

It is useful to distinguish two sources of a belief's intrinsic value: *valence* and *clarity.* Valence refers to the value attached to answers to questions. To illustrate the concept of valence, we return to the example of a professor's belief about her teaching ability. Being a good (or bad) teacher carries intrinsically positive (or, respectively, negative) valence. Clarity refers to preferences between degrees of certainty, independent of the answers one is certain of. We assume that, *ceteris paribus, people prefer to have greater clarity (i.e., less uncertainty or more definitive subjective beliefs).* The aversion that people feel towards uncertainty is reflected in neural responses in the anterior cingulate cortex, the insula and the amygdala (Hirsh and Inzlicht, 2008; Sarinopoulos et al., 2010). It manifests in physiological responses as well. Subjects who know to expect an electric shock, but who are uncertain whether it will be mild or intense, show more fear – they sweat more profusely, and their hearts beat faster – than subjects who know for sure that an intense shock awaits (Arntz et al., 1992).

When valence and clarity pull in opposite directions, it may be the case that people prefer a certain answer to a subjective belief that dominates it on valence or that people prefer uncertainty when it leaves space for better answers. While the preference for clarity violates Savage's (1954) sure-thing principle, we do assume a weaker version of it:

One-Sided Sure-Thing Principle

For any $\pi \in \Delta(\alpha)$, let $\text{supp}(\pi) \subseteq \alpha$ denote the support of π. If for all $\mathbf{A} \times x \in \text{supp}(\pi)$ we have $u(\pi', \mathbf{w}) \geq u(\pi^{\mathbf{A}\times x}, \mathbf{w})$, then $u(\pi', \mathbf{w}) \geq u(\pi, \mathbf{w})$, with the latter inequality strict whenever there exist $\mathbf{A}' \times x'$ and $\mathbf{A}'' \times x'' \in \text{supp}(\pi)$ such that $\mathbf{A}' \neq \mathbf{A}''$.

The one-sided sure-thing principle asserts that people always prefer a certain answer to uncertainty amongst answers that all have valences no better than the certain answer (holding attention weight constant).

A Measure of Uncertainty

The assumption of a preference for clarity means that there is a preference for less uncertain subjective beliefs. A useful measure of the uncertainty about a particular question is the entropy of the subjective probability distribution over answers (Shannon, 1948). The entropy of a subjective (marginal) probability π_i is $H(\pi_i) = -\sum_{A_i \in \mathcal{A}_i} \pi_i(A_i) \log \pi_i(A_i)$ (with the convention that $0 \log 0 = 0$).[13] At one extreme, entropy is high when there are many equally likely possible answers; at the other extreme, there is minimal entropy of 0 when a single answer is known for sure.

3.3 A Specific Utility Function

To make precise predictions about preferences for (or to avoid) information, we consider a specific utility function incorporating the preference for clarity and the role of attention weights:

$$u(\pi, \mathbf{w}) = \sum_{x \in X} \pi_X(x) v_X(x) + \sum_{i=1}^{m} w_i \left(\sum_{A_i \in \mathcal{A}_i} \pi_i(A_i) v_i(A_i) \; - H(\pi_i) \right). \quad (2)$$

We represent the value of prize x as $v_X(x)$ and the valence of answer A_i as $v_i(A_i)$. We now describe properties (some quite strong and almost certainly not always satisfied) that characterize (and necessarily imply) this utility function (see Theorem 1 below).

4. CHARACTERIZATION OF THE UTILITY FUNCTION

4.1 Properties

The utility function in Equation (2) satisfies the following seven properties.

Independence Across Prizes

In Section 2 we assumed independence across cognitive states. Independence might extend, as in traditional models, to material outcomes, holding beliefs constant.

P1. Holding the rest of the cognitive state constant, the preference relation satisfies independence across prizes if $u(\pi^{\mathbf{A}}, \mathbf{w}) = \sum_{x \in X} \pi_X^{\mathbf{A}}(x)\, u(\pi^{\mathbf{A}\times x}, \mathbf{w})$.

Property (P1) implies belief-dependent expected utility over lotteries that are independent of beliefs about the world. If

[13] The base of the logarithm in the entropy formula is arbitrary and amounts to a normalization parameter.

we also were to assume belief-independent utility for prizes, then we would gain the ability to reduce compound lotteries consisting of horse races as well as roulette lotteries (Anscombe and Aumann, 1963) to single-stage lotteries. However, we believe it is often the case that utility is belief-dependent. We might say that a decision maker often has a horse in the race.

Separability Between Questions

Additive separability of utility between questions means that a person can place a value on a belief about a given question without needing to consider beliefs about other questions.

P2. A utility function satisfies additive separability between questions if $u(\pi, \mathbf{w}) = u_X(\pi_X) + \sum_{i=1}^{m} u_i(\pi_i, w_i)$.[14]

Property (P2) may seem quite strong because we can imagine representations of sensible preferences that are not additively separable. For example, the value of a belief about whether a car on sale has a warranty intuitively could depend on the cost of the car in the first place (not to mention one's desire for a new car, one's estimation of the costs of car repairs, etc.). However, we may be able to represent these preferences as separable after all. We might suppose that these beliefs do have separable values but that they correlate with some other highly valued belief, perhaps about how good a deal one can get on the car. That is, while intuition tells us that the value of beliefs about different questions (e.g., "does she like me?" and "does she have a boyfriend?") is often interdependent, this dependence may be mediated by the existence of additional questions (e.g., "will she go out with me?"), beliefs about which may be mutually dependent, but independently valued.

Monotonicity with respect to Attention Weights

Preferences satisfy the property of monotonicity with respect to attention weights if whenever increasing attention on a given belief enhances (or diminishes) utility, it will do so regardless of the absolute level of attention weight. At a psychological level, the interpretation of this monotonicity property is that when a belief is positive, more attention to it is always better, and when a belief is negative, more attention is always worse. In fact, the property provides a natural *definition* of whether a belief is positive or negative.

P3. Preferences satisfy monotonicity with respect to attention weights if for any $\mathbf{w}$, $\hat{\mathbf{w}}$, and $\hat{\hat{\mathbf{w}}} \in \mathbb{R}_+^m$ such that $w_i = \hat{w}_i = \hat{\hat{w}}_i$ for all $i \neq j$ and $\hat{\hat{w}}_j > \hat{w}_j > w_j$, we have $u(\pi, \hat{\mathbf{w}}) \geq u(\pi, \mathbf{w})$ if and only if $u(\pi, \hat{\hat{\mathbf{w}}}) \geq u(\pi, \hat{\mathbf{w}})$, with equality on one side implying equality on the other, for all $\pi \in \Delta(\alpha)$.

In the case that these inequalities hold strictly, we say that π_j, the belief about question Q_j, is a *positive belief.* If they hold as equalities, we say that π_j is a *neutral belief.* And, in the case that the inequalities hold in the reverse direction, then π_j is a *negative belief.*

Linearity with respect to Attention Weights

The next property describes how changing the attention on a belief impacts utility. For any given attention weight, the marginal utility of a change in belief depends on what those beliefs are and how much the individual values them. The property of linearity with respect to attention weights means that, in general, the marginal utility associated with such a change in belief (assuming the utility of this belief is separable) is proportional to the attention on that belief.

P4. When the utility of question Q_i is separable, linearity with respect to attention weights is satisfied if for any w_i and $\hat{w}_i \in \mathbb{R}_+$ and π_i' and $\pi_i'' \in \Delta(\mathcal{A}_i)$, we have

$$u_i(\pi_i', \hat{w}_i) - u_i(\pi_i'', \hat{w}_i) = \frac{\hat{w}_i}{w_i} \left(u_i(\pi_i', w_i) - u_i(\pi_i'', w_i) \right).$$

Property (P4) allows us, in the case of separable utility, to assign an intrinsic value v to beliefs such that $u_i(\pi_i', w_i) - u_i(\pi_i'', w_i) = w_i \left(v_i(\pi_i') - v_i(\pi_i'') \right)$. We abuse notation by referring to the valence of answer A_i as $v_i(A_i)$, with it being defined here as the intrinsic value v_i of belief with certainty in A_i. We have taken the liberty of specifying a precise relationship between attention weights and utility as a convenient simplification; it should be noncontroversial because we do not claim to have a cardinal measure of attention weight.

Label Independence

Intuitively, the value of a belief should depend on how an individual values the possible answers and on how probable each of these answers is, and these factors (controlling for attention weight of course) should be sufficient to determine the utility of any (uncertain) belief. In particular, the value of a belief should not depend on how the question or the answers are labeled.

P5. Label independence is satisfied if, when the utility of questions Q_i and Q_j are separable, a bijection $\tau : \mathcal{A}_i \to \mathcal{A}_j$, such that $v_i(A_i) = v_j(\tau(A_i))$ and $\pi_i(A_i) = \pi_j(\tau(A_i))$, implies that $v_i(\pi_i) = v_j(\pi_j)$.

Reduction of Compound Questions

The intuition behind the assumption of label independence also seems to suggest that the utility of a belief perhaps should not depend on the way the question giving rise to the belief is asked, i.e., on whether a complicated question is broken up into pieces. We should recall, however, that the activation of a particular question directs attention to the belief about this question. Thus, in general, the utility of a belief will not be invariant to the question being asked. Still, it may be the case that utility remains invariant when a compound question is broken into parts as long as the attention on each part is weighted properly. If utility remains invariant upon setting attention weights on conditional questions to be proportional to the subjective probabilities of the hypothetical conditions, then we say that the utility function satisfies the reduction of compound questions property. Figure 1 demonstrates the reduction of a compound question with appropriate attention weights on each subquestion.

P6. A separable utility function satisfies the reduction of compound questions property if whenever there is a partition ζ of the answers $\mathcal{A}_i$ (to question Q_i) into $\zeta = \{\mathcal{A}_{i_1}, \ldots, \mathcal{A}_{i_n}\}$

[14] A subset of questions $\tilde{\mathcal{Q}} \subset \mathcal{Q}$ can also be separable, in which case $u(\pi, \mathbf{w}) = \sum_{i: Q_i \in \tilde{\mathcal{Q}}} u_i(\pi_i, w_i) + u_{-\tilde{\mathcal{Q}}}(\pi_{-\tilde{\mathcal{Q}}}, \mathbf{w}_{-\tilde{\mathcal{Q}}})$ where $\pi_{-\tilde{\mathcal{Q}}}$ is the marginal distribution over answers to the remaining questions and prizes and the vector $\mathbf{w}_{-\tilde{\mathcal{Q}}}$ contains the remaining components of $\mathbf{w}$.

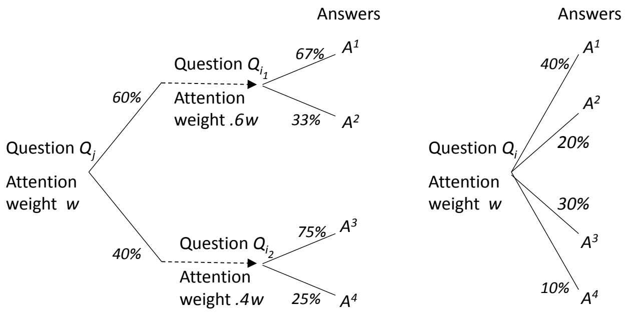


Figure 1: Decomposition of a compound question.

and a bijection $\tau : \zeta \to \mathcal{A}_j$ into the answers to some question Q_j such that for any $h \in [1, n]$ and any $A_i \in \mathcal{A}_{i_h}$,

$$v_i(A_i) = v_j(\tau(\mathcal{A}_{i_h})) + v_{i_h}(A_i)$$

and

$$\pi_i(A_i) = \pi_j(\tau(\mathcal{A}_{i_h})) \cdot \pi_{i_h}(A_i),$$

it follows that

$$u_i(\pi_i, \omega) = u_j(\pi_j, \omega) + \sum_{h=1}^{n} u_{i_h}(\pi_{i_h}, \pi_j(\tau(\mathcal{A}_{i_h})) \cdot \omega).$$

Ruling Out Unlikely Answers Increases Clarity

A final property operationalizes the preference for clarity. Controlling for the valence of one's beliefs, by considering situations in which one is indifferent between different possible answers to a question, there should be a universal aversion to being uncertain about the answer to an activated question. As a building block toward quantifying the uncertainty in a subjective belief, we assert here that when an unlikely (and equally attractive) answer is ruled out, uncertainty decreases (and thus the utility of that uncertain belief increases).

P7. Ruling out unlikely answers increases clarity if, when the utility of question Q_i is separable and all answers to this question have the same valence, i.e. $v_i(A_i) = v_i(A_i')$ for all A_i and $A_i' \in \mathcal{A}_i$, then for any π where without loss of generality $\pi_i(A_i^h)$ is weakly decreasing in h and for any π' such that $\pi_i'(A_i^h) \geq \pi_i(A_i^h)$ for all $h \in [1, \bar{h}]$ (with at least one inequality strict) and $\pi_i'(A_i^h) = 0$ for all $h > \bar{h}$, for some $\bar{h}$, we consequently have $v_i(\pi_i') > v_i(\pi_i)$.

4.2 Utility Representation Theorem

THEOREM 1. *If the properties P1-P7 are satisfied, then*

$$u(\pi, \mathbf{w}) = \sum_{x \in X} \pi_X(x) v_X(x) + \sum_{i=1}^{m} w_i \left(\sum_{A_i \in \mathcal{A}_i} \pi_i(A_i) v_i(A_i) \; - H(\pi_i) \right).$$

PROOF. Linearity with respect to attention weights allows us to pull an attention weight on question Q_i outside of the utility $u_i(\pi_i, w_i) = w_i v_i(\pi_i)$ (using a neutral belief to calibrate v_i). A partition of $\mathcal{A}_i$ into singletons $\mathcal{A}_{i_h}$ such that $v_i(A_i) = v_{i_h}(A_i)$ allows us, by reduction of the compound question, to determine that the function $F(\pi_i) = v_i(\pi_i) - \sum_{A_i \in \mathcal{A}_i} \pi_i(A_i) v_i(A_i)$ does not depend on $v_i(A_i)$ for any $A_i \in \mathcal{A}_i$. Moreover, $-F(\cdot)$ satisfies Shannon's (1948) axioms (continuity, increasing in the number of equiprobable answers, and reduction of compound questions) characterizing the entropy function $H(\pi_i) = -\sum_{A_i \in \mathcal{A}_i} \pi_i(A_i) \log \pi_i(A_i)$. □

5. INFORMATION ACQUISITION AND AVOIDANCE

We can apply our utility function to decisions about information acquisition or avoidance. We develop our analysis in a companion paper (Golman and Loewenstein, 2015a), and we provide a broad outline here of its implications. The desire for information, in our model, can be decomposed into three distinct motives: recognition of the instrumental value of the information; curiosity to fill the information gap(s); and motivated attention to think more or less about what could be discovered. The instrumental value of information arises from its impact on subsequent actions. As in the standard account of informational preferences, it is defined as the difference between the expected utility of subsequent actions conditional on having the information and the utility expected in the absence of the information. Curiosity arises from the expected reduction in uncertainty upon acquiring information. It is defined as the expected utility of revised beliefs, given prior levels of attention. The magnitude of curiosity depends on the attention devoted to each information gap that stands to be addressed. Motivated attention arises from the surprise upon acquiring information. It is defined as the expected utility from increased attention on whatever happens to be discovered, conditioning on all possible outcomes. Motivated attention is a motive to acquire information that's expected to be good and to avoid information that's expected to be bad.

Putting the three motives together, our model makes many predictions about when, and the degree to which, information will be sought or avoided. When anticipated answers are neutral or even potentially positive, information should be sought. The strength of the desire for this information should increase with the number of attention gaps that can be addressed, the attention paid to them, and the valence of the possible outcomes. However, when anticipated outcomes are sufficiently negative, information would be avoided. This "ostrich effect" when anticipating bad outcomes is consistent with a growing body of empirical evidence (see, e.g., Karlsson et al., 2009; Eil and Rao, 2011). In addition, the belief-resolution effect in our model leads to a novel prediction: individuals who discount the future less should be less likely to exhibit the ostrich effect and more likely to acquire information despite anticipated bad news.

6. RISK AND AMBIGUITY PREFERENCE

Section 5 outlines how the model we have developed allows us to describe a desire to acquire or to avoid information. We can apply this same model to an entirely new domain: preferences about wagers that depend on missing information. Risk and ambiguity aversion are complex topics, and we develop these applications in depth in a companion paper (Golman and Loewenstein, 2015b). Here, we provide a broad outline of the model's implications in this domain.

Decision making under risk and under ambiguity both expose decision makers to information gaps. Imagine a choice between a gamble and a sure thing. Deciding to play the gamble naturally focuses attention on the question: what will be the outcome of the gamble? Of course, deciding to not play the gamble does not stop an individual from paying some attention to the same question (or, if not choosing the gamble means it will not be played out, the related question: what *would have been* the outcome of the gamble?) but playing the gamble makes the question more important, and that brings about an increase in the attention weight on the question. If the individual is aware of this effect, which is natural to assume, then whether it encourages risk taking or risk aversion will depend on a second factor: whether thinking about the information gap is pleasurable or aversive. When thinking about the missing information is pleasurable, then the individual will be motivated to increase attention on the question, which entails betting on it. Conversely, when thinking about the missing information is aversive, the individual will prefer to not bet on it. This may help to explain why, for example, people generally prefer to bet on their home teams rather than on other teams, especially in comparison to the home team's opponent.

Decision making involving uncertainties that are ambiguous is similar to the case with known risks, but with an additional wrinkle: with ambiguity, there are additional information gaps. In a choice between a sure thing and an ambiguous gamble, for example, a second relevant question (in addition to the one above about the outcome of the gamble) is: what is the probability of winning with the ambiguous gamble? (And there may be additional relevant questions that could inform someone about this probability, so even a Bayesian capable of making subjective probability judgments would be exposed to these information gaps.) Again, betting on the ambiguous gamble makes these questions more important and thus will increase the attention weight on them. So, desire to play the gamble will be increasing with the degree to which thinking about the gamble is pleasurable. To the extent that abstract uncertainties are not pleasurable to think about, this model provides a novel account of standard demonstrations of ambiguity aversion, including those first generated by Ellsberg (1961) in his seminal paper on the topic.

7. DESIRE FOR WISDOM

Our utility model can be used to describe preferences between knowing and not knowing. But another comparison is also of interest, albeit harder to investigate empirically – the difference between awareness and unawareness. While we cannot easily give a person the choice whether or not to become aware of a question, we can at least introspect. We might posit that awareness of meaningful questions is a source of utility. Equation (2), the utility function which represents preferences between cognitive states given a fixed set of activated questions $\mathcal{Q}$, might be augmented with a term $v_{\mathcal{Q}}(\mathcal{Q})$ capturing the intrinsic value of awareness of particular issues.

Wisdom, the combination of awareness and clarity,[15] is, or at least tends to be, preferable to ignorance. We of course must allow exceptions if we are serious that beliefs have valence that may be negative. The popular adage that "ignorance is bliss" expresses concern for the negative beliefs that awareness may entail. However, in many natural situations, a person may reasonably anticipate that newfound awareness will bring about neutral or even positive beliefs. In such contexts, information and awareness may be simultaneously acquired. For example, a bird-watcher typically would strictly prefer to learn the name of a previously unnoticed songbird rather than to remain unaware of its existence. Curiosity is behind the desire to catch the name upon becoming aware of the bird's existence, even though the particular name does not really matter, but utility from awareness implies that opening, and then immediately closing, an aversive information gap need not be zero sum. Rather, discovering the new bird's name, acquiring both the question and the definitive answer, produces a net positive utility gain, which is what we designate, in the context of our model, the *utility of wisdom.* We find the desire for wisdom in individuals' varied pursuits of insight and expertise, from a naturalist's passion for identifying flora and fauna to a fan's thirst for new baseball statistics or a connoisseur's discriminating taste for wine.[16]

Question	Answer	Belief		
Latent	–	Unawareness	↓ Awareness	↓Wisdom
Activated	Unknown	Uncertainty	↓ Clarity	
	Known	Certainty		

Table 4: Wisdom, the combination of awareness and clarity.

Aristotle in 350 B.C. asserted, "All men by nature desire to know." John Stuart Mill agreed, in his classic *Utilitarianism*, arguing that, "It is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied." We too assert that knowledge can be a very real source of utility. A perspective that information derives value solely from its ability to yield material consumption fails to appreciate the most profound benefits provided by information, the knowledge and wisdom it confers.

8. REFERENCES

[1] Abelson, R. (1986). Beliefs are Like Possessions. *Journal for the Theory of Social Behavior* 16 (3), 223-250.

[2] Anscombe, F., Aumann, R. (1963). A Definition of Subjective Probability, *Annals of Mathematical Statistics* 34, 199-205.

[3] Arntz, A., Van Eck, M., de Jong, P. (1992). Unpredictable Sudden Increases in Intensity of Pain and Acquired Fear. *Journal of Psychophysiology* 6 (1), 54-64.

[4] Asch, D., Patton, J., Hershey, J. (1990). Knowing for the Sake of Knowing: The Value of Prognostic Information. *Medical Decision Making* 10 (1), 47-57.

[5] Aumann, R. (1976). Agreeing to Disagree. *Annals of Statistics* 4, 1236-1239.

[6] Baldi, P. (2002). A Computational Theory of Surprise. In *Information, Coding, and Mathematics*,

[15] We are aware that this may not be the most common usage of the word *wisdom*, but the distinction between knowledge acquired from a state of uncertainty and knowledge acquired from a state of unawareness is rarely made explicit. The term, "wisdom" seems to adequately capture this distinction if we think of a wise man or woman as not only having the right answers, but also asking the right questions.

[16] Lab studies also find that people prefer environments which seem to stimulate new questions and promise to provide relevant information (Kaplan, 1992).

M. Blaum, P. Farrell, H. van Tilborg (Eds.) Norwell, MA: Kluwer.

[7] Caplin, A., Leahy, J. (2001). Psychological Expected Utility Theory And Anticipatory Feelings. *Quarterly Journal of Economics* 116 (1) 55-79.

[8] Cover, T., Thomas, J. (1991). *Elements of Information Theory.* New York: Wiley.

[9] Dekel, E., Lipman, B., Rustichini, A. (1998). Standard State-Space Models Preclude Unawareness. *Econometrica* 66 (1), 159-173.

[10] Dillenberger, D. (2010). Preferences for One-Shot Resolution of Uncertainty and Allais-Type Behavior. *Econometrica* 78 (6), 1973-2004.

[11] Eil, D., Rao, J. (2011). The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *American Economic Journal: Microeconomics* 3, 114-138.

[12] Eliaz, K., Spiegler, R. (2006). Can Anticipatory Feelings Explain Anomalous Choices of Information Sources? *Games and Economic Behavior* 56 (1), 87-104.

[13] Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, 75, 643-699.

[14] Fehr, E., Rangel, A. (2011). Neuroeconomic Foundations of Economic Choice – Recent Advances. *Journal of Economic Perspectives* 25 (4), 3-30.

[15] Gabaix, X., Laibson, D., Moloche, G., Weinberg, S. (2006). Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model. *American Economic Review* 96 (4), 1043-1068.

[16] Geanakoplos, J., Pearce, D., Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior* 1, 60-79.

[17] Golman, R., Loewenstein, G. (2015a). Curiosity, Information Gaps, and the Utility of Knowledge. Working Paper.

[18] Golman, R., Loewenstein, G. (2015b). Information Gaps for Risk and Ambiguity. Working Paper.

[19] Grant, S., Kajii, A., Polak, B. (1998). Intrinsic Preference for Information. *Journal of Economic Theory* 83, 233-259.

[20] Heifetz, A., Meier, M., Schipper, B. (2006). Interactive Unawareness. *Journal of Economic Theory* 130 (1), 78-94.

[21] Hirsh, J., Inzlicht, M. (2008). The Devil You Know: Neuroticism Predicts Neural Response to Uncertainty. *Psychological Science* 19 (10), 962-967.

[22] Hirshleifer, J., Riley, J. (1979). The Analytics of Uncertainty and Information – An Expository Survey. *Journal of Economic Literature* 17 (4), 1375-1421.

[23] Itti, L., Baldi, P. (2009). Bayesian Surprise Attracts Human Attention. *Vision Research* 49 (10), 1295-1306.

[24] Janis, I. (1958). *Psychological Stress: Psychoanalytic and Behavioral Studies of Surgical Patients.* New York: Wiley.

[25] Kadane, J., Schervish, M., Seidenfeld, T. (2008). Is Ignorance Bliss? *Journal of Philosophy* 105 (1), 5-36.

[26] Kang, M.J., Hsu, M., Krajbich, I., Loewenstein, G., McClure, S., Wang, J., Camerer, C. (2009). The Wick in the Candle of Learning: Epistemic Curiosity Activates Reward Circuitry and Enhances Memory. *Psychological Science* 20 (8), 963-973.

[27] Kaplan, S. (1992). Environmental preference in a knowledge-seeking, knowledge-using organism. In *The Adapted Mind*, J. H. Barkow, L. Cosmides, and J. Tooby (Eds.), New York: Oxford UP.

[28] Karlsson, N., Loewenstein, G., McCafferty, J. (2004). The Economics of Meaning. *Nordic Journal of Political Economy* 30 (1), 61-75.

[29] Karlsson, N., Loewenstein, G., Seppi, D. (2009). The Ostrich Effect: Selective Attention to Information. *Journal of Risk and Uncertainty* 38 (2), 95-115.

[30] Karni, Edi, and Marie-Louise Vierø. (2013). "Reverse Bayesianism": A Choice-Based Theory of Growing Awareness. *American Economic Review* 103 (7), 2790-2810.

[31] Klibanoff, P., Marinacci, M., Mukerji, S. (2005). A Smooth Model of Decision Making under Ambiguity. *Econometrica* 73, 1849-1892.

[32] Köszegi, B. (2010). Utility from Anticipation and Personal Equilibrium. *Economic Theory* 44 (3), 415-444.

[33] Kreps, D., Porteus, E. (1978). Temporal Resolution of Uncertainty and Dynamic Choice Theory. *Econometrica* 46 (1), 185-200.

[34] Li, J. (2008). A Note on Unawareness and Zero Probability. PIER Working Paper No. 08-022.

[35] Locke, E., Latham, G. (1990). *A Theory of Goal Setting and Task Performance.* Englewood Cliffs, NJ: Prentice Hall.

[36] Loewenstein, G. (1994). The Psychology of curiosity: A review and reinterpretation. *Psychological Bulletin* 116 (1), 75-98.

[37] Loewenstein, G. (1999). Because it is there: The challenge of mountaineering... for utility theory. *Kyklos* 52, 315-344.

[38] Modica, S., Rustichini, A. (1994). Awareness and Partitional Information Structures. *Theory and Decision* 37 (1), 107-124.

[39] Modica, S., Rustichini, A. (1999). Unawareness and Partitional Information Structures. *Games and Economic Behavior* 27 (2), 265-298.

[40] Miller, G., Galanter, E., Pribram, K.H. (1960). *Plans and the Structure of Behavior.* New York: Holt.

[41] Nielsen, L. (1984). Unbounded Expected Utility and Continuity. *Mathematical Social Sciences* 8, 201-216.

[42] Rabin, M. (2000). Risk Aversion and Expected-utility Theory: A Calibration Theorem. *Econometrica* 68, 1281-92.

[43] Ritov, I., Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making* 3 (4), 263-277.

[44] Sarinopoulos, I., Grupe, D., Mackiewicz, K., Herrington, J., Lor, M., Steege, E., Nitschke, J. (2010). Uncertainty During Anticipation Modulates Neural Responses to Aversion in Human Insula and Amygdala. *Cerebral Cortex* 20 (4), 929-940.

[45] Savage, L. (1954). *The Foundations of Statistics.* New York: Wiley.

[46] Schelling, T. (1987). The Mind as a Consuming Organ. In *The Multiple Self*, J. Elster (Ed.), Cambridge: Cambridge UP, 177-196.

[47] Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* 27, 379-423.

[48] Tasoff, J., Madarász, K. (2009). A Model of Attention and Anticipation. Working Paper.

[49] von Neumann, J., Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton UP.

[50] Wakker, P. (1988). Nonexpected Utility as Aversion of Information. *Journal of Behavioral Decision Making* 1 (3), 169-175.

[51] Yariv, L. (2001). Believe and Let Believe: Axiomatic Foundations for Belief Dependent Utility Functionals. Cowles Foundation Discussion Paper No. 1344.

Deliberating between Backward and Forward Induction Reasoning: First Steps

Aleks Knoks
University of Maryland
aknoks@umd.edu

Eric Pacuit
University of Maryland
epacuit@umd.edu

ABSTRACT

Backward and forward induction can be viewed as two styles of reasoning in dynamic games. Since each prescribes taking a different attitude towards the past moves of the other player(s), the strategies they identify as rational are sometimes incompatible. Our goal is to study players who are able to deliberate between backward and forward induction, as well as conditions under which one is superior to the other. This extended abstract is our first step towards this goal. We present an extension of Stalnaker's game models [34, 35], in which the players can make "trembling hand" mistakes. This means that when a player observes an unexpected move, she has to figure out whether it is a result of a deliberate choice or a mistake, thereby committing herself to one of the two styles of reasoning.

1. INTRODUCTION AND MOTIVATION

We begin with a motivating example. Consider the game G_1 depicted in Figure 1. There are two players: Ann (A) and Bob (B). Ann moves first (node h_0) and can either choose to go out (O), immediately ending the game, or stay in the game (I). If she chooses to stay in, node h_1 is reached. At h_1, Ann and Bob move simultaneously (Ann's available actions are u and v while Bob's are a, b and c). The structure of this game is similar to the extensively studied *Battle of the Sexes with an Outside Option* (see, for instance, [7, 14, 37]).

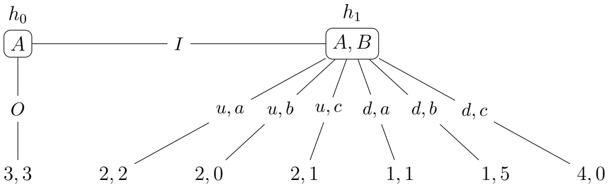


Figure 1: The game G_1

Suppose that Bob initially believes that Ann is going to choose O. To his amazement, however, Ann stays in the game. Now Bob has to figure out why Ann decided to play I, what she will choose at node h_1, and, most importantly, what is his own best response. There are two plausible "lines of reasoning" for Bob. The first goes as follows: Ann chose I at h_0 because she is hoping for a payoff greater than 3. Thus, Ann must be hoping that the game will terminate in the rightmost node. Hence, the rational choice for Bob is b, which—assuming his conjecture about Ann is correct—would result in a payoff of 5. Alternatively, Bob can avoid speculating about the reasons behind Ann's move and focus on trying to figure out what is the rational thing for her to do at h_1. Although Ann might hope initially that the game will end in the rightmost node, she must realize that Bob will never choose c since it is strictly dominated by a. So, guaranteeing a payoff of 2, u is Ann's rational choice at h_1. Clearly, if Bob convinces himself that Ann is choosing u, playing a is his best response. The two lines of reasoning, thus, lead to different recommendations for Bob. The first is an example of the so-called "forward induction reasoning" requiring that the players think critically about the observed past choices of their opponent(s) and find plausible explanations for these choices [7, 24, 28, 35, 37]. The second can be called "backward induction reasoning" requiring the players to only reason about their opponents' future behavior and not about their past moves [2, 12, 28, 30, 35].

There are many characterizations of both forward and backward induction reasoning in the game theory literature (cf. [7, 28, 35]). These formal renderings match the informal explanation given above, recommending that Bob plays a and b, respectively. However, the formal models do not solve what we take to be Bob's real challenge, namely, deciding which of these two lines of reasoning is more plausible in the present case.[1] Notice that a wrong choice leads to an unwelcome consequence. Suppose that Bob interprets Ann's choice of I as an attempt to get a higher payoff, but it turns out that she did it for some other reason—e.g. she was careless—and that, at h_1, she decides to play u. In this case, Bob ends up with 0. Now, suppose that Bob disregards Ann's previous move, as backward induction suggests he should, but it turns out that Ann is hoping to get 4. In this case, Bob's payoff is 1 instead of 5.

These considerations bring us to the following general question: How can a player deliberate between backward and forward induction in cases in which both seem plausible (at least *prima facie*) while dictating incompatible choices? Admittedly, in the above situation, Bob seems to be faced with a particularly difficult choice, since his information does not seem to sway the scales in favor of one or the other style of reasoning. But suppose that he is in a situation in which the players are prone to making *mistakes* relatively frequently—we elaborate on this notion below; for now simply think of

[1] We do not mean to suggest that Bob is explicitly applying backward or forward induction himself. Rather, a theorist can identify his reasoning as an instance of one or the other. In our models, the players' "choice" of reasoning style will be traced back to their prior beliefs about how likely it is that their opponents may make a mistake.

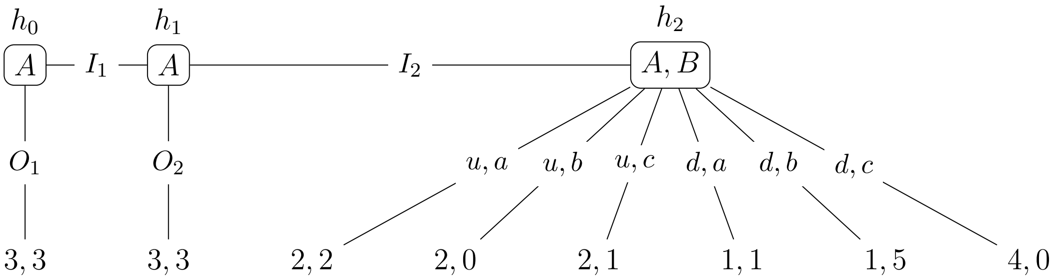


Figure 2: The game G_2

the so-called "trembling hand mistakes" [32]: Ann chooses O, but plays I instead. In this case, backward induction reasoning should be preferred. Or let's say that Ann and Bob are playing a different game, G_2, in which mistakes are still possible, but Ann has two opportunities to go out before reaching the node at which the players move simultaneously—see Figure 2. Intuitively, if Bob observes Ann play I_1, it is reasonable for him to interpret her move as a mistake. Suppose, however, that Bob subsequently observes that Ann plays I_2. Now, the interpretation of her previous choices that is suggested by forward induction seems more plausible (we could of course modify the game further by adding more opportunities for Ann to exit the game). What this all suggests is that additional information about the *context of the game*—e.g., how probable it is that players make mistakes—can help the player settle on a style of reasoning.

Our ultimate goal is to study players that are able to deliberate between forward and backward induction, as well as conditions which make each style of reasoning superior to the other. The above considerations suggest that to do so we need to endow standard game theoretic agents with (relatively) rich beliefs. We use a model introduced by Stalnaker [33, 34, 35, 36] to describe the players' beliefs. In addition to having beliefs about their opponents' strategies and the game—standard for Epistemic Game Theory [16, 27, 29]—our players can also interpret observed behavior as either the result of deliberate action or as a mistake.

This paper is structured as follows. Section 2 describes the formal framework that allows us to represent the two lines of reasoning discussed above: extensive games with simultaneous moves (Section 2.1), our extension of Stalnaker's model (Section 2.2), two notions of rationality (Section 2.3), as well as an illustrative example (Section 2.4). In Section 3, we argue that our model offers an illuminating perspective on epistemic characterizations of backward induction and is a conservative extension of Stalnaker's model. Finally, Sections 4 and 5 discuss related work and outline a number of directions for future research.

2. FRAMEWORK

2.1 Extensive games with simultaneous moves

The examples from the introduction are **extensive games with simultaneous moves**.[2] Following [26, Section 6.3.2], we describe them as structures $\langle N, Act, H, \tau, \{u_i\}_{i \in N}\rangle$, where:

- N is a finite set of players.

- Act is the set of actions available to the players. To simplify notation, we assume that Act is partitioned into sets of actions for each player. For player $i \in N$, let $Act_i \subseteq Act$ denote player i's actions.

- H is a set of finite sequences of finite sequences of elements of Act. Elements $h \in H$ are called **histories**. We assume H satisfies the following constraints:

  - $\epsilon \in H$, where ϵ denotes the empty history.
  - If $h \in H$ and $h' \preceq h$, then $h' \in H$, where $h' \preceq h$ means that h' is a initial segment of h. Formally, we write $h' \preceq h$ provided $h = h'u$ where u is a sequence of sequences from Act, and $h'u$ denotes the concatenation of h' with u.
  - Each $h \in H$ is finite. That is, we restrict attention to *finite horizon* games.[3] We write $len(h)$ for the **length** of h (i.e., the number of elements in h).

  A history $h \in H$ is called a **terminal history** if there is no $h' \in H$ such that $h' \neq \epsilon$ and $hh' \in H$. Let $Z \subseteq H$ denote the set of terminal histories. Let $V = H - Z$ be the set of non-terminal histories. Each non-terminal history is associated with a simultaneous decision problem for a set of players. For this reason, we sometimes call elements of V **decision nodes**. For $h \in V$, let $A(h)$ be the possible extensions of h:

  $$A(h) = \{\vec{a} \mid h\vec{a} \in H \text{ and } \vec{a} \text{ is a sequence of actions}\}.$$

- τ is a turn function $\tau : V \to \wp(N)$ assigning a set of players[4] to each non-terminal history $h \in V$. For each $i \in N$, let $V_i = \{h \in V \mid i \in \tau(v)\}$ be the set of non-terminal histories where player i moves. Similarly, we define the set of actions available to i at a decision node $h \in V_i$. For each $h \in V$ and $i \in \tau(h)$, let $A_i(h)$ be the set of actions available to i at h:

  $$A_i(h) = \{a \in Act_i \mid \text{there is an } \vec{a} \in A(h) \text{ containing } a\}$$

  If $i \notin \tau(h)$, then let $A_i(h) = \emptyset$. We impose an additional constraint to ensure that each decision node is

[2] We assume that the reader is familiar with the basics of game theory. The formal definitions are included here to fix notation.

[3] This is a standard restriction in the literature on epistemic characterization of backward induction.

[4] Throughout this article, we assume that for all $h \in V$, $\tau(h) \neq \emptyset$. If we drop this assumption, then histories in which $\tau(h) = \emptyset$ should be interpreted as a move by nature.

associated with a *strategic game*:[5]

- For each $h \in V$, $A(h) = \Pi_{i \in \tau(h)} A_i(v)$.

- For each $i \in N$, $u_i : Z \to \mathbb{R}$ is a utility function.

A **strategy** for player i assigns an action to each of i's decision nodes. Formally, a strategy for player i is a function $s_i : V_i \to Act$ where for all $h \in V_i$, $s_i(h) \in A_i(h)$. Let S_i be the set of all strategies for player i. As usual, a **strategy profile** is a sequence of strategies, one for each player (i.e., an element of $\Pi_{i \in N} S_i$). Given a strategy profile $\mathbf{s}$, let $\mathbf{s}_i$ be player i's component of $\mathbf{s}$ and $\mathbf{s}_{-i}$ the sequence of strategies from $\mathbf{s}$ for all players except i (i.e., $\mathbf{s}_{-i} \in \Pi_{j \neq i} S_j$). Each profile of strategies $\mathbf{s}$ generates a terminal history $\rho_{\mathbf{s}} \in Z$. We say that a non-terminal history is **reached** by a strategy profile provided h is an initial segment of $\rho_{\mathbf{s}}$.

A strategy for player i represents her *conditional plan* for the game. It prescribes a choice for player i at all of i's decision nodes, including those that are *ruled-out* by the strategy itself. Suppose that $h \in V_i$. An action $a \in Act_i(h)$ **rules out** a decision node $h' \in V_i$ provided $h \preceq h'$, but $h\vec{a} \not\preceq h'$ for any $\vec{a} \in A(h)$ containing a. In addition, we say that a rules out action a' provided $a' \in A_i(h')$ for some decision node $h' \in V_i$ that is ruled-out by a.[6] For example, in Figure 1, A's action O at h_0 rules out the actions u and d (because O rules out h_1).

2.2 Game models

A **game model** describes the players' beliefs during a play of the game. As discussed in the introduction, we are interested in representing players that allow for the possibility that one or more of their opponents made a *mistake*. This means that we must include states in which the moves of player i (i.e., the observed *behavior* of player i) does not match i's *choices*. To make this precise, each state in the game model will be associated with both strategies for the players *and* sequences of actions representing the observed behavior of the players.

The players' *behavior* in a game is represented by a sequence of actions. Recall that histories $h \in H$ are sequences of sequences of actions (one action for each player whose turn it is to move). For each $h \in H$, let $beh_i(h)$ be the sequence of i's actions in h. Formally, beh_i is defined by induction on the length of histories: $beh_i(\epsilon) = \epsilon$ (at the initial node, none of the players have made a choice), and

$$beh_i(h\vec{a}) = \begin{cases} beh_i(h)a & \text{if } i \in \tau(h) \text{ and } \vec{a} \text{ contains } a \\ beh_i(h) & i \notin \tau(h) \end{cases}$$

If X is a set, then X^* is the set of all finite strings of X. An i-history is a sequence of actions such that $\alpha = beh_i(h)$ for some $h \in H$. Given an i-history α and a decision node $h \in V_i$, let α_h be the component of α describing the action chosen at h. If α does not specify a move at h (either because the previous moves in α rule out h or α is not a maximal history), then α_h is undefined. For instance, in Figure 1, there are four A-histories (ϵ, O, Iu, and Id) and four B-histories (ϵ, a, b, and c). We use Ou to denote the strategy s_A in which $s_A(h_0) = O$ and $s_A(h_1) = u$ (similarly, for Od). Furthermore, we have $Iu_{h_0} = I$, $Iu_{h_1} = u$, and O_{h_1} is undefined.

A player history may be a *partial* description of what that player does in the game. This happens when the i-history α does not specify a choice for i at a decision node h not ruled out by α. Of course, if an i-history α specifies an action for player i at a decision node $h \in V_i$, then α specifies an action for i at each h' such that $h' \preceq h$ and $h' \in V_i$. We are interested in sets of player histories that represent possible plays of the game. A set of player histories $\{\alpha_i\}_{i \in N}$ is **coherent** if there is a history $h \in H$ such that for all $i \in N$, $beh_i(h) = \alpha_i$. Note that a set of i-histories may be coherent, yet not completely describe a path trough the game. For instance, $\{I, c\}$ is a coherent set of player histories in the game pictured in Figure 1: There are two histories $h = (I)(u, c)$ and $h' = (I)(d, c)$ such that $beh_B(h) = beh_B(h') = c$ and $beh_A(h) = beh_A(h') = I$. However, there is a unique history representing the play of the game associated with a coherent set of player strategies. The play of the game generated by a coherent set of i-histories $\{\alpha_i\}_{i \in N}$ is the longest history h such that $h \preceq h'$ for each h' such that for all $i \in N$, $beh_i(h') = \alpha_i$. The play of the game associated with the coherent set $\{I, c\}$ in the game in Figure 1 is $h = (I)$. The play of the game associated with a coherent set of player histories may be empty and need not be maximal. For example, the following table lists the coherent sets of strategies and the corresponding play of the game for the game pictured in Figure 1.

Coherent sets player strategies	Play of the game
$\{\epsilon, \epsilon\}$	ϵ
$\{O, \epsilon\}$	(O)
$\{I, a\}$, $\{I, b\}$, $\{I, c\}$, $\{I, \epsilon\}$	(I)
$\{Iu, \epsilon\}$, $\{Id, \epsilon\}$	(I)
$\{Iu, a\}$	$(I)(u, a)$
$\{Iu, b\}$	$(I)(u, b)$
$\{Iu, c\}$	$(I)(u, c)$
$\{Id, a\}$	$(I)(d, a)$
$\{Id, b\}$	$(I)(d, b)$
$\{Id, c\}$	$(I)(d, c)$

The sets $\{O, a\}$, $\{O, b\}$ and $\{O, c\}$ are not coherent.

Suppose that W is a nonempty set, elements of which are called **states**. Each player i will be associated with two functions β_i and σ_i subject to the following constraints:

1. For each $i \in N$, $\beta_i(w)$ is a (possibly empty) i-history and $\sigma_i(w)$ is a strategy for player i.

2. The i-histories $\{\beta_i(w)\}_{i \in N}$ are **coherent**.

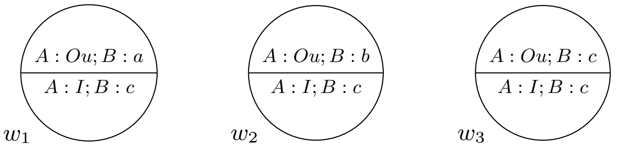
We say that a player made a **mistake at a history** $h \in V_i$ in the world w provided her behavior is different than what

[5] There is a hidden notational difficulty here. Since different players move at different decision nodes, the indices of the sequences of actions change from decision node to decision node. Formally, we represent a sequence $\vec{a}$ at decision node h as a function $\vec{a} : \tau(h) \to \cup_{i \in \tau(h)} A_i(h)$ where for each $i \in \tau(h)$, $\vec{a}(h) \in A_i(h)$. We write $\vec{a}_i$ to denote the action $a \in A_i(h)$ such that $\vec{a}(i) = a$ and say $\vec{a}$ contains a. This implicitly assumes that $i \in \tau(h)$ (otherwise $\vec{a}_i$ is not well-defined). Alternatively, we could assume that all players move at every decision node and introduce notation to distinguish "active" players from "passive" players. The passive players at a decision node would only have a single action available for their choice. We follow the first approach in this article.

[6] We are implicitly assuming that all the action labels are unique. This assumption can be dropped, although it does simplify the notation.

is prescribed by her chosen strategy $\sigma_i(w)$ at h. Formally, i made a mistake at $h \in V_i$ provided $\beta_i(w)_h \neq \sigma_i(w)(h)$ (if $\beta_i(w)_h$ is defined).

Example. Recall the game in Figure 1 and consider three states w_1, w_2 and w_3. Suppose that $\sigma_A(w_1) = \sigma_A(w_2) = \sigma_A(w_3) = Ou$ (recall that Ou is the strategy in which A chooses O at h_0 and u at h_1, but it is not an A-history) and $\beta_A(w_1) = \beta_A(w_2) = \beta_A(w_3) = I$. Thus, A made a mistake at h_0. The strategies for player B are $\sigma_B(w_1) = a$, $\sigma_B(w_2) = b$, $\sigma_B(w_3) = c$ (again, these are the strategies in which B chooses, respectively, a, b, and c at h_1), and $\beta_B(w_1) = \beta_B(w_2) = \beta_B(w_3) = c$. These states are pictured as follows:



The strategies $\sigma_A(w)$ and $\sigma_B(w)$ are displayed in the top half of the circles and the histories $\beta_A(w)$ and $\beta_B(w)$ in the bottom half (where $w \in \{w_1, w_2, w_3\}$). If these states describe A's beliefs (i.e., they are the set of doxastic possibilities for A), then A is certain that B will play c, but is uncertain about exactly *why* B is playing c. It might be because B made a mistake (as in states w_1 and w_2) or because B simply followed through on his plan to play c. Furthermore, A arrived at these beliefs under the supposition that she (contrary to her chosen strategy) selected I at h_0.

The players' beliefs and belief revision policies are represented in the standard way (cf. [4, 10, 36]). Each player $i \in N$ is associated with a **prior probability** on the set of states, $P_i \in \Delta(W)$,[7] and a **plausibility ordering** $\succeq_i \subseteq W \times W$ satisfying the following constraints: for each $i \in N$ and for each $w \in W$, $P_i(w) > 0$ (i.e., P_i is a full support probability measure); $\succeq_i$ is a locally connected (for all w, v, x, if $w \succeq_i x$ and $w \succeq_i v$, then $v \succeq_i x$ or $x \succeq_i v$) partial order (reflexive and transitive relation) on W. The plausibility ordering $\succeq_i$ represents player i's belief revision policy. For each $i \in N$ and states $w, v \in W$, let $w \approx_i v$ iff $w \succeq_i v$ or $v \succeq_i w$. Since, $\succeq_i$ is a locally complete partial order, $\approx_i$ is an equivalence relation. For $w \in W$, let $[w]_i = \{v \mid w \approx_i v\}$ denote the equivalence class of w for $\approx_i$, called i's *information cell.* The intended interpretation is that $w \approx_i v$ means that w and v are *subjectively indistinguishable* to player i (i's beliefs, knowledge, and conditional beliefs are the same in both states).[8] The players' *full beliefs* at a state w are defined as usual: For each $w \in W$, let $\max_{\succeq_i}([w]_i) = \{x \mid \text{there is no } y \in [w]_i \text{ such that } y \succ_i x\}$, where $y \succ_i x$ means that $y \succeq_i x$ but $x \not\succeq_i y$.

Definition 1. For each $w \in W$ and $i \in N$, player i's **(partial) beliefs** at state w are given by the probability measure $P_{i,w} \in \Delta(W)$ defined as follows: For each $E \subseteq W$,

$$P_{i,w}(E) = P_i(E \mid \max_{\succeq_i}([w]_i))$$

The players' partial beliefs $P_{i,w}$ represent their beliefs about the possible choices, behaviors and beliefs of their opponents at state w.[9] The belief revision policy describes how the players revise their beliefs given *any* evidence $F \subseteq W$:

$$P_{i,w}(E \mid F) = P_i(E \mid \max_{\succeq_i}(F \cap [w]_i)).$$

Note that this conditional probability is well-defined for any set F such that $F \cap [w]_i \neq \emptyset$. In particular, there may be a set F such that $P_{i,w}(F) = 0$, yet $P_{i,w}(\cdot \mid F)$ is well-defined. This is a very general model of belief revision for the players, since it describes how each player revises her beliefs given *any* evidence consistent with her current information (i.e., any F such that $[w]_i \cap F \neq \emptyset$). However, we are primarily interested in how the players revise their beliefs given the actions that they observe in the game.[10] Each state $w \in W$ is associated with a history $h \in H$ as follows. Let h_w be the history corresponding to the play of game associated with $\{\beta_i(w)\}_{i \in N}$ (see the discussion above). Note that h_w need not be a maximal history, so h_w is the behavior that is observed at state w. For any $h \in H$, let $[h] = \{w \mid \beta_i(w) = beh_i(h) \text{ for all } i \in N\}$ be the event that the players behaved according to history h. Then, $P_{i,w}(E \mid [h_w])$ is i's probability of E given her most plausible explanation of the actions she observed at state w. Thus, the belief revision policy describes how the players' beliefs change during a play of the game.[11]

Putting everything together, a **game model** for a game G is a tuple $\mathcal{M}_G = \langle W, \{(\beta_i, \sigma_i)\}_{i \in N}, \{\succeq_i\}_{i \in N}, \{P_i\}_{i \in N}\rangle$. In addition, we impose the following two constraints:

- For all $w \in W$ and $i \in N$, if $v \in [w]_i$, then $\sigma_i(w) = \sigma_i(v)$. That is, players *know* their own strategy.[12]
- For all $w \in W$ and $i \in N$, for each initial segment $h' \subseteq h_w$ (including the empty history), there is a $w' \in [w]_i$ such that $h_w = h'$.

The last constraint ensures that if a sequence of choices in the game is consistent with a player's information, then all of its initial segments must be consistent with the player's information. This is a consequence of assuming that the structure of the game is (commonly) known to all the players and that players cannot think it is possible to observe a history without observing the sequence of choices that generated the history. Compare the above constraint with the stronger assumption that for all $w \in W$, for all $h \in H$, there is a $w' \in [w]_i$ such that $h_{w'} = h$. This ensures that it is consistent with the players' information that every possible history in the game could be realized. Of course, *ex ante*, the players do not rule out any histories.[13] However, our models represent the players' *ex iterim* beliefs. In such models, it may be

[7] For a set X, let $\Delta(X)$ be the set of all probability measures on X. In this paper, we assume that the set of states is finite, so we can assume that P is defined on all subsets of W.

[8] That is, the equivalence classes of $\approx_i$ are the different "types" for player i.

[9] That is, beliefs about the possible types of their opponents.

[10] Our models of games are closely related to Bayesian extensive games with observable actions [26, Section 12.3]. However, there are important methodological and conceptual differences between Bayesian games and epistemic models of games (see [27, Section 1.4]). For this reason, we postpone a complete comparison between our game models and Bayesian extensive games with observable actions to the full version of the paper.

[11] Thus, our models are related to the *type spaces* based on conditional probability systems from [7, Section 2.2].

[12] Each player can be associated with a standard knowledge operator where for all E, $K_i(E) = \{w \mid [w]_i \subseteq E\}$.

[13] Assuming that all the players are *aware* (in the sense of [22, 23]) of the structure of the game.

consistent with a player's information (which includes her chosen strategy) that some history of the game will not be played (cf. the discussion of richness conditions on the model in Section 5).

2.3 Rationality

A player chooses rationally provided her strategy choice at a state maximizes the players subjective expected utility with respect to her beliefs about the past and expected moves of her opponents. We do not assess the rationality of the players' moves themselves. Thus, a player may choose rationally at a state, though she may not carry out her plan because she made a mistake.

Suppose that G is an extensive game with simultaneous moves and $\mathcal{M}_G = \langle W, \{(\beta_i, \sigma_i)\}_{i\in N}, \{\succeq_i\}_{i\in N}, \{P_i\}_{i\in N}\rangle$ is a game model for G. For each $w \in W$, the strategy realized at w by player i is $s_i(w) : V_i \to Act_i$ defined as follows:

$$s_i(w)(h) = \begin{cases} \beta_i(w)_h & \text{if } \beta_i(w)_h \text{ is defined} \\ \sigma_i(w)(h) & \text{otherwise} \end{cases}$$

Then, $\mathbf{s}(w) = (s_1(w), \ldots, s_n(w))$ is a profile of strategies, and let $Out(\mathbf{s})$ be the (unique) terminal history generated by $\mathbf{s}$.

Definition 2. For any strategy $s_i \in S_i$ for player i, the **expected utility** of s_i at state w is:

$$EU_{i,w}(s_i) = \sum_{w'\in W} P_{i,w}(\{w'\} \mid [h_w])u_i(Out(s_i, \mathbf{s}_{-i}(w))).$$

A player chooses optimally at state w provided her current strategy maximizes her subjective expected utility at w, given the actions that she observed. Let $S_i(w) \subseteq S_i$ be the set of strategies for player i that conform to player i's moves in state w. That is, $s_i \in S_i(w)$ implies that for all $h \in V_i$, if $\beta_i(w)_h$ is defined, then $s_i(h) = \beta_i(w)_h$. Then,

$$\begin{aligned} Opt_i &= \{w \mid \sigma_i(w) \text{ maximizes expected utility} \\ &\quad \text{with respect to } P_{i,w} \text{ and } S_i(w)\}. \end{aligned}$$

If $w \in Opt_i$, then player i is adopting the best possible strategy given i's observations at w. Rationality is more demanding. There are two versions of rationality. The first requires that a player is rational at a state w provided her strategy at w is optimal given her beliefs at w *and* was optimal at all previous decision nodes given her beliefs at the moment of decision. We say that a state $w' \in [w]_i$ is an **earlier choice state** provided $\beta_i(w')$ is an initial segment of $\beta_i(w)$.

Definition 3. Player i is **rational-1** at state w provided $w' \in Opt_i$ for *all* earlier choice states w'. Let Rat_i^1 be the set of all states w such that i is rational-1 in w.

A player may be rational-1 even if she does not correctly implement her strategy. The second version of rationality requires that a player's strategy is optimal even when the player learns that her beliefs are mistaken. That is, the strategy is optimal and remains optimal after any belief revision.

Definition 4. Player i is **rational-2** at state w provided $w' \in Opt_i$ for *all* states $w' \in [w]_i$. I.e., $[w]_i \subseteq Opt_i$. Let Rat_i^2 be the set of all states w such that i is rational-2 in w.

Of course, $Rat_i^1 \subseteq Rat_i^2$ (if a player is rational-1, then the player is rational-2). However, in general, the converse is not true (this is illustrated by an example in the next section).

2.4 Example

Figure 3 depicts models of the games from Figures 1 and 2. These models represent the players' initial beliefs and dispositions to change their beliefs that we discussed in the introduction. The model on the left, $\mathcal{M}_1$, represents one play of the game in Figure 1, and the model on the right, $\mathcal{M}_2$, represents a play of the game in Figure 2. We draw an arrow from state v to state w when $w \succeq_i v$. The solid arrows represent Bob's plausibility ordering $\succeq_B$ and the dashed arrows represent Ann's plausibility ordering $\succeq_A$ (we only represents Bob's beliefs in $\mathcal{M}_2$). To keep down the clutter in the pictures, we assume that the remaining arrows can be inferred by transitivity and reflexivity. The strategies $\sigma_A(w)$ and $\sigma_B(w)$ are displayed in the top half of the circles and the histories $\beta_A(w)$ and $\beta_B(w)$ in the bottom half (empty histories are left blank). We think of the players strategy choices and moves as discrete random variables. Thus, $[Choose_i^h = a] = \{w \mid \sigma_i(w)(h) = a\}$ is the event that player i chooses action a at decision node h. Similarly, $[Move_i^h = a] = \{w \mid \beta_i(w)_h = a\}$ is the event that player i played a at history h. The (common) prior probabilities are displayed next to the states.

Suppose that w_4 is the actual world in model $\mathcal{M}_1$. Thus, Ann chose the strategy Ou, but made a mistake and played I followed by u (as originally planned). Bob chose strategy a which he correctly implemented when given the chance to move. His (overall) most plausible worlds are w_1 and w_2. This means that he is certain that Ann plays O at h_0 (i.e., $P_{B,w_4}([Choose_A^{h_0} = O]) = 1$). Moreover, he (initially) thinks that Ann's strategies Ou and Od are equally likely (i.e., $P_{B,w_4}([Choose_A^{h_1} = u]) = P_{B,w_4}([Choose_A^{h_1} = d]) = 0.5$). If Ann surprises Bob by playing I, he is disposed to interpret this as a mistake on her part, rather than as revealing that she is following a different strategy (i.e., $\max_{\succeq_B}([w_4]_B \cap [Move_A^{h_0} = I]) = \{w_3\}$, $\beta_A(w_3) = I$ while $\sigma_A(w_3) = Ou$). Furthermore, after observing Ann play I, Bob is certain that her next move will be u: $P_{B,w_4}([Choose_A^{h_1} = u] \mid [Move_A^{h_0} = I]) = 1$. This model also illustrates what it means for a player to be rational-1. Note that Ann made a mistake in w_4, yet she is still rational-1 ($w_4 \in Rat_A^1$). Both w_1 and w_3 are earlier choice states for Ann (as is w_4), and she chooses optimally in all these states: $Opt_A = \{w_1, w_2, w_3\}$.

The model $\mathcal{M}_2$ in Figure 3 represents Bob's beliefs in the game from Figure 2 in which Ann has two opportunities to exit the game. Suppose that w_6 is the actual world. No mistakes are made with Ann playing I_1I_2c and Bob playing u. Initially, Bob believes that Ann is going to choose O_1 ($\max_{\succeq_B}([w_6]_B) = \{w_1, w_2\}$ with $\sigma_A(w_1)(h_0) = \sigma_A(w_2)(h_0) = O_1$). On the condition that Ann actually plays I_1, he is disposed to interpret her move as a mistake, predicting that she is going to go out at the next opportunity ($\max_{\succeq_B}([w_6]_B \cap [Move_A^{h_1} = I_1]) = \{w_3, w_4\}$, $\sigma_A(w_3)(h_1) = \sigma_A(w_4)(h_1) = O_2$. If Ann surprises Bob the second time by playing I_2, he is disposed to conclude that it is very likely that Ann actually chose to play I_1 and I_2 ($P_{B,w_6}([Choose_A^{h_0} = I_1] \cap [Choose_A^{h_1} = I_2] \mid [Move_A^{h_0} = I_1] \cap [Move_A^{h_1} = I_2]) = .9$). Intuitively, if Ann surprises Bob, he is disposed to reason in the backward induction style (ignoring her mistake), but if she surprises him a second time, Bob switches to forward induction and conjectures that (it is highly probable that) Ann is going to play d. The model $\mathcal{M}_2$ also illustrates the difference between rationality-1 and rationality-2. In w_2, Bob

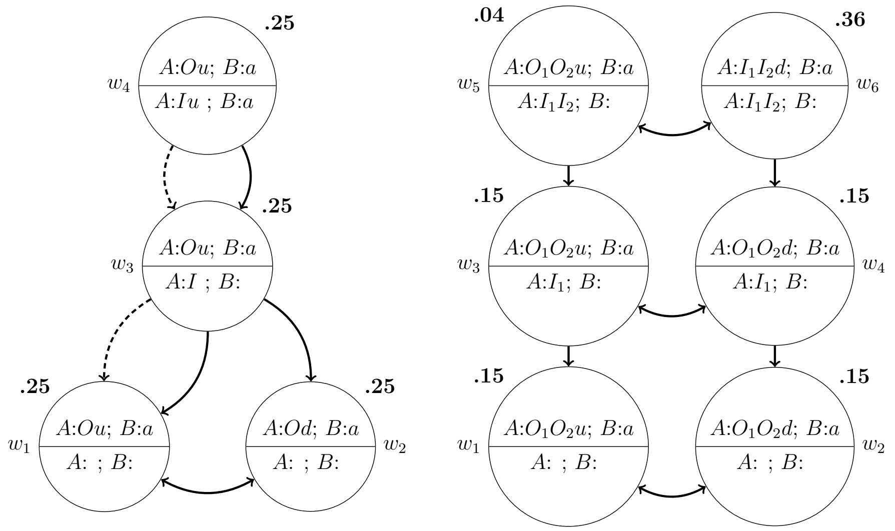


Figure 3: The models $\mathcal{M}_1$ and $\mathcal{M}_2$

is rational-1, since $\{w_1, w_2\} \subseteq Opt_B$, but he is not rational-2, since $w_6 \in [w_2]_B$, but $w_6 \notin Opt_B$. That is, the strategy that Bob chooses at w_2 is not optimal *with respect to the beliefs he would have* after revising his initial beliefs with the information that Ann plays I_1 and I_2.

3. STALNAKER AND AUMANN

It is easy to see that our game model is a conservative extension of Stalnaker's [34, 35, 36]. Since we extend his model by allowing states in which the players' moves differ from what is prescribed by their chosen strategy, the players can *know* each other's strategies and still be uncertain about the way the game is going to end. In spite of this, however, our models can accommodate the standard epistemic characterizations of backward induction found in the literature, and, in particular, Aumann's classic characterization.

Aumann proved that if there is common knowledge that all of the players are rational, then the backward induction path will be realized [2]. Our models are richer than Aumann's: We describe the players' beliefs and belief revision policies in addition to the players' knowledge. Recently, Samet extended Aumann's result to doxastic models, which are much closer to the models we use. He proved that if there is *common belief*[14] that all the players' strategies are *doxastically substantively rational*, then the backward induction path is realized [31]. Since we allow for mistakes, we will have models in which there is common belief that the players choose optimally, but the backward induction path does not obtain. There is another difference between our models and Aumann's and Samet's. The behavior functions can be viewed as a temporal parameter. That is, our model includes states that describe the players' beliefs at different moments during the play of the game (cf. [5, 12]). In general, the players' beliefs may change even if the game unfolds according to their chosen strategies. We can recover Samet's characterization of backward induction with an additional constraint:

For all $i \in N$ and $w, w' \in W$, if $w' \in [w]_i$, then for each $w'' \in \max_{\succeq_i}([w]_i \cap [h_{w'}])$ there is a $w''' \in \max_{\succeq_i}([w]_i)$ such that $\sigma_i(w'') = \sigma_i(w''')$ for all $i \in N$.

This constraint says that the players cannot learn anything about their opponents' strategies that they did not already know at the beginning of the game.

Proposition 1. *Suppose that G is an extensive game (without simultaneous moves) in "general position" (see Appendix A) and $\mathcal{M}_G$ is a model for G satisfying the above constraint and that every possible mistake is considered: for all $w \in W$, every possible mistake that i can make given i's strategy at w is realized by the behavior at some state $w' \in [w]_i$. Suppose that $w \in W$ is a state in which the histories $(\beta_1(w), \dots, \beta_n(w))$ generate a maximal path through the game. If there are no mistakes in w and common belief at w that all the players are rational-1, then the path that is generated by the histories is the backward induction path.*

There are many other epistemic characterizations of backward induction.[15] What is more relevant for our purposes is Stalnaker's criticism of Aumann's epistemic characterization of backward induction [35, Section 5]. The problem lies with Aumann's notion of rationality which is captured and refined by our rationality-1. A player is rational-1 provided her strategy is optimal (given a sequence of moves) and was optimal at all previous choices *with respect to her beliefs at the moment of choice.* Stalnaker argues that this notion of rationality is much too strong.[16] His idea is that a strategy

[14] We assume that the reader is familiar with the formal definition of common belief. See Appendix A for the formal definition in our framework.

[15] It is beyond the scope of this article to survey all of the different approaches. See [29, Section 8.11] and [27] for a discussion and pointers to the literature.

[16] We will not repeat Stalnaker's argument here. The gist of it is that it is important not to conflate "action a would be optimal if node v were reached" and "if node v *is* reached, then action a is optimal".

for player i is optimal provided i would choose optimally at node v (according to his strategy) given i's beliefs *under the hypothesis that node v is reached*. To formalize this idea, Stalnaker introduces the notion of *perfect rationality*: "In cases where two or more [strategies] are [optimal], the agent should consider, in choosing between them, how he should act if he learned he was in error about something." [34, pg. 148]. It is not hard to see that our definition of rationality-2 is equivalent to Stalnaker's definition of perfect rationality. Thus, our models can accommodate both Stalnaker's and Aumann's analysis of backward induction. Of course, this, by itself, is not new (cf. [5] and [19]). However, our analysis also opens the door for further refinements of the notions of rationality they use.

For instance, perfect rationality (or rationality-2) requires that a player's strategy is *robustly optimal*. That is, it is optimal even after the player learns that her beliefs are mistaken. Variants of rationality-2 can be defined by fixing the set of *evidence* that may induce a change of belief for a player. For instance, we can require that a player's strategy must be robustly optimal with respect to evidence about her opponents' moves (cf. [7, Section 2.2]), strategy choices, beliefs, or even evidence about the player's own moves. A complete analysis of the different options will be left for the full version of the paper.

4. RELATED WORK

Our model allows for states in which the players' moves differ from what is prescribed by their chosen strategy. This general idea (i.e., trembling hand mistakes) was used by Selten and others to characterize *refinements* of the Nash equilibrium (cf. [18, 32]). Within the equilibrium refinement program, Bicchieri's work on forward and backward induction [9] comes closest to ours. In [9], the players respond to (hypothetical) surprising moves in an extensive game (that may be the result of a trembling-hand mistake) by revising their beliefs à la AGM [1]. Our models differ in both important technical details and the underlying motivations. Most importantly, we downplay the role that the Nash equilibrium (and its refinements) plays in the analysis of rational behavior in game situations (this is in line with much of the epistemic game theory literature, cf. [13]).

More recently, Cubitt and Sugden develop a model in which a player's behavior may, in principle, differ from her (rational) choice [15]. They include a postulate stating that the players' behavior must all conform to the same principles of rational choice. Among other things, they are interested in highlighting the role that this assumption plays in the players' *reasoning* about what to do in a game situation (cf. also Bacharach's discussion of the *transparency of reason* in [3, Section 4.2]). There are some intriguing connections between our work and theirs, but a complete discussion will be left for the full version of this paper.

5. CONCLUDING REMARKS

We have imposed only two minimal constraints on our models: every information cell must include the player's beliefs at all previous choice points, and the players "know" their own strategy choice. The literature on forward induction, and, more generally, belief revision in games [5, 11, 35], contains other natural constraints that we may want to impose. One belief revision policy that has been extensively discussed in relation to forward induction reasoning is the so-called **rationalizability principle** [8]: "A player should always try to interpret her information about the behavior of her opponents assuming that they are not implementing 'irrational' strategies." (cf. [6]). In order to represent this belief revision policy, Stalnaker includes a "richness" condition on his models [35, pg. 35, footnote 5] ensuring that the players have the conditional beliefs needed to rationalize any observed behavior.[17] With such a richness condition, we can formally prove Stalnaker's characterization of the belief revision policy in which the players apply the rationalizability principle at most once.[18]

Another direction for future research is to compare our approach to belief revision with non-standard probabilities, lexicographic probability systems, and conditional probability systems [21, 25]. Once the relationship between these different models is understood, we can connect our work with Battgalli and Siniscalchi's characterizations of common *strong belief* of rationality [7] and Halpern's recent epistemic characterizations of trembling-hand equilibria using non-standard probabilities [20].

Finally, note that the games in Figures 1 and 2 have the same reduced normal form. However, our analysis in this paper suggests that there are strategically relevant differences between the two games (cf. [24]). In particular, the players may be able to *learn* about their opponents' strategies during a play of the game. This suggests possible connections with models of learning in extensive games [17].

6. REFERENCES

[1] C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510 – 530, 1985.

[2] R. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8(1):6 – 19, 1995.

[3] M. Bacharach. A theory of rational decision in games. *Erkenntnis*, 27(1):17 – 55, 1987.

[4] A. Baltag and S. Smets. Conditioanl doxastic models: A qualitative approach to dynamic belief revision. In *Electornic notes in theoretical computer science*, volume 165, pages 5 – 21. Springer, 2006.

[5] A. Baltag, S. Smets, and J. Zvesper. Keep 'hoping' for rationality: a solution to the backwards induction paradox. *Synthese*, 169:301–333, 2009.

[6] P. Battigalli. On rationalizability in extensive games. *Journal of Economic Theory*, 74:40 – 61, 1997.

[7] P. Battigalli and M. Siniscalchi. Strong belief and forward induction reasoning. *Journal of Economic Theory*, 106(2):356 – 391, 2002.

[8] D. Bernheim. Rationalizable strategic behavior. *Econometrica*, 52:1007 – 1028, 1984.

[9] C. Bicchieri. *Rationality and Coordination (Cambridge Studies in Probability, Induction, and Decision*

[17] This is related to Battigalli and Siniscalchi's use of *complete* (conditional) type spaces in their characterization of extensive-form rationalizability and related concepts. See [7, Section 6.2.3] for a discussion.

[18] Stalnaker states and hints at a proof of such a theorem in [35, pg. 51, footnote 22].

Theory). Cambridge: Cambridge University Press, 1993.

[10] O. Board. Dynamic interactive epistemology. *Games and Economic Behavior*, 49(1):49–80, 2004.

[11] G. Bonanno. AGM belief revision in dynamic games. In K. R. Apt, editor, *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge (TARK XIII)*, pages 37 – 45, 2011.

[12] G. Bonanno. A doxastic behavioral characterization of generalized backward induction. *Games and Economic Behavior*, 88:221 – 241, 2014.

[13] A. Brandenburger. Introduction. In *The Language of Game Theory: Putting Epistemics into the Mathematics of Games*. World Scientific Series in Economic Theory, 2014.

[14] R. Cooper, D. DeJong, R. Forsythe, and T. Ross. Forward induction in the battle-of-the-sexes games. *The American Economic Review*, 83(5):1303 – 1316, 1993.

[15] R. P. Cubitt and R. Sugden. Common reasoning in games: A Lewisian analysis of common knowledge of rationality. *Economics and Philosophy*, 30:285 – 329, 2014.

[16] E. Dekel and M. Siniscalchi. Epistemic game theory. In H. P. Young and S. Zamir, editors, *Handbook of Game Theory, Volume 4*. Elsevier, 2014.

[17] D. Fudenberg and D. M. Kreps. Learning in extensive-form games I. self-confirming equilibria. *Games and Economic Behavior*, 8:20 – 55, 1995.

[18] S. Govindan and R. Wilson. Nash equilibrium, refinements of. In S. N. Durlauf and L. E. Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke, 2008.

[19] J. Halpern. Substantive rationality and backward induction. *Games and Economic Behavior*, 37(2):425 – 435, 2001.

[20] J. Halpern. A nonstandard characterization of sequential equilibrium, perfect equilibrium, and proper equilibrium. *International Journal of Game Theory*, 3838(1):37–50, 2009.

[21] J. Halpern. Lexicographic probability, conditional probability, and nonstandard probability. *Games and Economic Behavior*, 68(1):155 – 179, 2010.

[22] J. Halpern and L. Rêgo. Extensive games with possibly unaware players. *Mathematical Social Sciences*, 70:42 – 58, 2014.

[23] A. Heifetz, M. Meier, and B. Schipper. Interactive unawareness. *Journal of Economic Theory*, 130(1):78 – 94, 2006.

[24] E. Kohllberg and J. Francois Mertens. On the strategic stability of equilibria. *Econometrica*, 54(5):1003 – 1037, 1986.

[25] D. Lehman and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55(1):1 – 60, 1992.

[26] M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

[27] E. Pacuit and O. Roy. Epistemic foundations of games. In E. N. Zalta, editor, *Stanford Encyclopedia of Philosophy*, 2015.

[28] A. Perea. Backward induction versus forward induction reasoning. *Games*, 1(3):168–188, 2010.

[29] A. Perea. *Epistemic Game Theory: Reasoning and Choice*. Cambridge UP, 2012.

[30] A. Perea. Belief in the opponents' future rationality. *Games and Economic Behavior*, 83:231 – 254, 2014.

[31] D. Samet. Common belief of rationality in games of perfect information. *Games and Economic Behavior*, 79:192 – 200, 2013.

[32] R. Selten. A reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4:25 – 55, 1975.

[33] R. Stalnaker. On the evaluation of solution concepts. *Theory and Decision*, 37(1):49 – 73, 1994.

[34] R. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12(02):133 – 163, 1996.

[35] R. Stalnaker. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36(1):31 – 56, 1998.

[36] R. Stalnaker. Extensive and strategic forms: Games and models for games. *Research in Economics*, 53(3):293 – 319, 1999.

[37] E. van Damme. Stable equilibria and forward induction. *Journal of Economic Theory*, 48:476 – 496, 1989.

APPENDIX

A. PROOF OF PROPOSITION 1

In this appendix, we restrict attention to extensive games $G = \langle N, Act, H, \tau, \{u_i\}_{i\in N}\rangle$ without simultaneous moves. So, for all decision nodes $v \in V$, $|\tau(v)| = 1$. In this case, we can view histories as sequences of actions rather than sequences of sequences of actions. Furthermore, following [2], we assume that the payoff for each of the players is different at different terminal nodes (the game is in "general position"). This implies that the result of applying the backward induction algorithm[19] is uniquely defined.

Belief operators: Suppose that

$$\mathcal{M}_G = \langle W, \{(\beta_i, \sigma_i)\}_{i\in N}, \{\succeq_i\}_{i\in N}, \{P_i\}_{i\in N}\rangle$$

is a game model. For each event $E \subseteq W$, we say that player i believes E, $B_i(E)$, provided E is implied by i's full beliefs. That is, $B_i(E) = \{w \mid \max_{\succeq_i}([w]_i) \subseteq E\}$.

Samet's game model: Samet's game model is a tuple $\langle W, \{\Pi_i, t_i\}_{i\in N}, \mathbf{s}\rangle$, where W is a non-empty set of states, for each $i \in N$, Π_i is a partition on W, $t_i : W \to \Delta(W)$ is a type function assigning a probability measure to each state, and $\mathbf{s} : W \to S$ (where $S = \Pi_i S_i$) assigns a strategy to each state. Let $[\mathbf{s}_i(w) = s_i]$ be the set of states w such that $\mathbf{s}_i(w) = s_i$. The knowledge and belief operators are defined as usual: for all $E \subseteq W$, $K_i(E) = \{w \mid \Pi_i(w) \subseteq E\}$ and $B_i(E) = \{w \mid t_i(w)(E) = 1\}$. Samet includes the following constraints:

[19] The terminal nodes are labeled with the payoffs for each player. For each non-terminal history h with $\tau(h) = \{i\}$, label h with the maximum of all the labels of the successors of h. This labeling is then used to identify the so-called *backward induction path*.

- For all $w \in W$, if $v \in \Pi_i(w)$, then $t_i(w) = t_i(v)$
- For all $w \in W$, $t_i(w)(\Pi_i(w)) = 1$
- For all $w \in W$, if $v \in \Pi_i(w)$, then $[\mathbf{s}_i(w) = s_i] \subseteq B_i([\mathbf{s}_i(w) = s_i])$.

Rationality in Samet's model: Building on the notation introduced in Section 2, for a strategy profile $\mathbf{s}$, let $Out_h(\mathbf{s})$ be the (unique) terminal history that is reached if the players follow their strategies in $\mathbf{s}$ starting at h. Then, for a state $w \in W$ and strategy $s_i \in S_i$, $Out_v(s_i, \mathbf{s}_{-i}(w))$ is the terminal node that is reached, starting at h, if player i follows the strategy s_i and the other players follow the strategies associated with state w. Then, let

$$[Out_h(s_i, \mathbf{s}_{-i}) >_i Out_h(\mathbf{s})] \quad = \{w \mid u_i(Out_h(s_i, \mathbf{s}_{-i}(w))) > u_i(Out_h(\mathbf{s}(w)))\}.$$

Player i is said to be **doxastically substantively rational** at all states when:

$$R_i^{ds} = \bigcap_{h \in V_i} \bigcap_{s_i \in S_i} \neg B_i([Out_h(s_i, \mathbf{s}_{-i}) >_i Out_h(\mathbf{s})])$$

Let $R^{ds} = \bigcap_{i \in N} R_i^{ds}$.

Common belief: Given belief operators $B_i : \wp(W) \to \wp(W)$ for each player $i \in N$ (defined in Samet's game model or our game model), we define a common belief operator $CB : \wp(W) \to \wp(W)$ in the usual way. First, define *everyone believes*: For all $E \subseteq W$, $B(E) = \bigcap_{i \in N} B_i(E)$. Then define the nth power of B, B^n, as follows: for all $E \subseteq W$, $B^1(E) = B(E)$ and for $n > 1$, $B^n(E) = B(B^{n-1}(E))$. Finally, **common belief** of an event E is $CB(E) = \bigcap_{n \geq 1} B^n(E)$

Samet's Theorem 3 states that, in any of his models, $CB(R^{ds}) \subseteq I$, where I is the set of states in which the backward induction path is played.

Suppose that $\mathcal{M}_G = \langle W, \{(\beta_i, \sigma_i)\}_{i \in N}, \{\succeq_i\}_{i \in N}, \{P_i\}_{i \in N} \rangle$ is a game model for our game G. The **forgetful projection of** $\mathcal{M}_G$, denoted $\mathcal{M}_G^\circ$, is the tuple $\langle W, \{\Pi_i, t_i\}_{i \in N}, \mathbf{s} \rangle$, where for each $w \in W$, let $\Pi_i(w) = [w]_i$, $t_i(w) = P_{i,w}$, and $\mathbf{s}(w) = (\sigma_1(w), \ldots, \sigma_n(w))$. It is not hard to see that $\mathcal{M}_G^\circ$ satisfies the constraints imposed by Samet. For instance, we have, for all $w' \in \Pi_i(w)$, $t_i(w) = t_i(w')$, since if $w' \approx_i w$, then $\max_{\succeq_i}([w]_i) = \max_{\succeq_i}([w']_i)$.

We first state and prove a simple Lemma that will be used to relate Samet's notion of doxastic substantive rationality with our rationality-1.

Lemma 1. *Suppose that the game G and game model $\mathcal{M}_G$ and state in w satisfy the assumption of Proposition 1. Then, for all players $i \in N$, for all $w' \in W$, if $w \in Rat_i^1$, then for all $v \in V_i$, there is some $w' \in \max_{\succeq_i}([w]_i)$ such that $u_i(Out_h(\mathbf{s}(w')) > u_i(Out_h(s_i; \mathbf{s}_{-i}(w')))$.*

Proof. First of all, it is easy to see that if a strategy s_i is optimal for player i at state w, then for all strategies $t_i \neq s_i$, there must be at least one state $w' \in \max_{\succeq_i}([w]_i \cap [Move_i^w = \beta_i(w))$ such that $u_i(Out(s_i; \mathbf{s}_{-i}(w))) > u_i(Out(t_i; \mathbf{s}_{-i}(w)))$.

Suppose that $w \in Rat_i^1$. Then, for all $h \in V_i$, if $\beta_i(w)_h$ is defined (i.e., h on the path generated by the behavior of the players in state w), then for all $s_i \in S_i(w)$, there is at least one state $w' \in \max_{\succeq_i}([w]_i \cap [h_w])$ such that $u_i(Out_h(\mathbf{s}(w))) > u_i(Out_h(s_i; \mathbf{s}_{-i}(w)))$. Note that we can move from $Out(\cdot)$ to $Out_h(\cdot)$ since we restrict attention to strategy profiles that conform to the behavior of the players at w. By the constraint stated before Proposition 1, this implies that there is a $w'' \in \max_{\succeq_i}([w]_i)$ such that $u_i(Out_h(\mathbf{s}(w''))) > u_i(Out_h(s_i; \mathbf{s}_{-i}(w'')))$. This, together with the assumption that all mistakes are realized by some state in i's information cell, ensures that, for every decision node $h \in V_i$, there is some $w' \in \max_{\succeq_i}([w]_i)$ such that

$$u_i(Out_h(\mathbf{s}(w'))) > u_i(Out_h(s_i; \mathbf{s}_{-i}(w'))).$$

This completes the proof of the Lemma. □

The proof of the proposition follows immediately:

Proof of Proposition 1. Suppose that $w \in W$ and $(\beta_1(w), \ldots, \beta_n(w))$ generate a maximal path through the game. If $w \in CB(\bigcap_j Rat_j^1)$ in $\mathcal{M}_G$, then Lemma 1 implies that $w \in CB(R^{ds})$ in the forgetful projection $\mathcal{M}_G^\circ$. Since $\mathcal{M}_G^\circ$ is a Samet model of a game, Samet's Theorem 3 implies that $w \in I$. Since no mistakes are made in w, this implies that $(\beta_1(w), \ldots, \beta_n(w))$ is the backward induction path. □

Standard State Space Models of Unawareness

[Extended Abstract][*]

Peter Fritz
Faculty of Philosophy, University of Oxford
Jesus College, Turl Street
Oxford OX1 3DW, UK
peter.fritz@philosophy.ox.ac.uk

Harvey Lederman
Department of Philosophy, New York University
5 Washington Place
New York, NY 10003, USA
hsl306@nyu.edu

ABSTRACT

The impossibility theorem of Dekel, Lipman and Rustichini [8] has been thought to demonstrate that standard state-space models cannot be used to represent unawareness. We first show that Dekel, Lipman and Rustichini do not establish this claim. We then distinguish three notions of awareness, and argue that although one of them may not be adequately modeled using standard state spaces, there is no reason to think that standard state spaces cannot provide models of the other two notions. In fact, standard space models of these forms of awareness are attractively simple. They allow us to prove completeness and decidability results with ease, to carry over standard techniques from decision theory, and to add propositional quantifiers straightforwardly.

Keywords

awareness, standard state space models, epistemic logic

1. INTRODUCTION

Dekel, Lipman and Rustichini [8], hereafter, "DLR", claim to show that "standard state space models preclude unawareness". Their claim has achieved the status of orthodoxy.[1] The first task of this paper is to clear the way for standard state space models of unawareness by showing that the formal result DLR present does not establish their headline conclusion. DLR informally motivate certain axioms concerning unawareness, but in their formal impossibility result, they rely on the claim that these axioms hold at all states in the model. As section 2 argues, the assumption that axioms hold at all states of the model is unwarranted; in fact, DLR themselves reject it. While DLR's formal results are valid, they are not sufficiently general to rule out standard state space models of unawareness. As we show, the impossibility results do not hold if one assumes only that DLR's explicit assumptions about unawareness hold at some "real" states, as opposed to at all states. Even strengthening those explicit assumptions considerably does not reinstate the results.

But this does not yet vindicate standard state space models of unawareness. Section 3 presents a novel impossibility result which uses widely shared assumptions about awareness. The new impossibility theorem relies on the assumption that an agent who is aware of a conjunction is aware of its conjuncts. If awareness satisfies this assumption, then standard state space models do in fact preclude unawareness. We then distinguish three notions of awareness, and suggest that two important ones do not satisfy this assumption, leaving open the possibility that they could be adequately modeled by standard state space models.

The remainder of the paper continues in a more positive vein. We describe a simple class of standard state space models which represent key features of awareness. In section 4, we establish completeness and decidability for the logic of these models. We also show that adding propositional quantifiers, a topic which has presented major difficulties for existing approaches to awareness, is straightforward in our standard state space models. In section 5, we present one way of implementing a choice-based approach to decision theory within these models, and show how non-trivial unawareness is consistent with speculative trade. Section 6 concludes.

2. DLR'S TRIVIALITY RESULT

2.1 Standard State Space Models

Standard state space models for the knowledge and awareness of a single agent can be understood as certain tuples $\langle \Omega, k, a\rangle$. Ω is required to be a set, called the set of *states*, from which a set of events is derived by taking an *event* to be a set of states. k and a are functions on events, which represent the agent's knowledge and awareness, respectively: k maps each event E to the event $k(E)$ of the agent knowing E; a similarly takes each E to the event $a(E)$ of the agent being aware of E.

Such models are straightforwardly used to interpret a formal language in which one can talk about knowledge and awareness. Let L be such a language built up from proposition letters $p, q, \dots$, using a unary negation operator $\neg$, a binary conjunction connective $\wedge$ and two unary operators K and A, respectively ascribing knowledge and awareness to the agent. Formulas of this language are interpreted relative to a model $M = \langle \Omega, k, a\rangle$ and a valuation function v which maps each proposition letter p to the event $v(p)$. The interpretation uses a function $[\![\cdot]\!]_{M,v}$ which maps each formula φ of L to the event expressed by φ in M, which can be understood as the set of states in which φ is true in M. To state the constraints on such a function let $-E = \Omega\backslash E$.

$$[\![p]\!]_{M,v} = v(p)$$

$$[\![\neg\varphi]\!]_{M,v} = -[\![\varphi]\!]_{M,v}$$

$$[\![\varphi \wedge \psi]\!]_{M,v} = [\![\varphi]\!]_{M,v} \cap [\![\psi]\!]_{M,v}$$

[*] A draft of the full version of this paper is available at `http://users.ox.ac.uk/~hert2388/`

[1] See, e.g., [53, p. 2], [54], [24, p. 78], [25, p. 305], [26, p. 101], [40, p. 220], [31, p. 2790], [39, pp. 977–978], [27, p. 2454], [28, p. 257], [15, p. 42], [58, p. 516], and [57].

$$[\![K\varphi]\!]_{M,v} = k([\![\varphi]\!]_{M,v})$$

$$[\![A\varphi]\!]_{M,v} = a([\![\varphi]\!]_{M,v})$$

The agent being unaware of something can of course be understood as it not being the case that she is aware of it. We therefore syntactically use $U\varphi$ as an abbreviation for $\neg A\varphi$. Similarly, we introduce the other connectives of classical propositional logic as abbreviations, using $\vee$ for disjunction, $\rightarrow$ for material implication, $\leftrightarrow$ for bi-implication, and $\top$ and $\bot$ for an arbitrary tautology and contradiction, respectively. On the semantic side, we adopt the convention of writing fg for the composition of functions f and g, which allows us two write, e.g., $k - a(E)$ instead of $k(-(a(E)))$.

In order to express general constraints on these models, we say that a formula φ is *valid on M* if $[\![\varphi]\!]_{M,v} = \Omega$ for each valuation function v; this can be understood as requiring φ to be true in every state of M according to every valuation function. In order to limit this constraint to a particular state $\omega \in \Omega$, we say that φ is *valid in ω* if $\omega \in [\![\varphi]\!]_{M,v}$ for each valuation function v.

These models count as "standard" in the sense of DLR. First, the events expressed by $A\varphi$ and $K\varphi$ are each a function of the event expressed by φ. (DLR call this "event-sufficiency".) Second, negation is interpreted as set-complement and conjunction as intersection, so that all tautologies of classical propositional logic, such as $p \vee \neg p$, are interpreted as the set of all states in every model. (DLR call this assumption "real states".)

2.2 DLR on Standard State Space Models

DLR introduce three constraints on awareness, which can be stated using the following three axioms:

Plausibility: $Up \rightarrow (\neg Kp \wedge \neg K\neg Kp)$

KU-Introspection: $\neg KUp$

AU-Introspection: $Up \rightarrow UUp$

Their constraints on knowledge can be stated using the following three axioms:

Necessitation: $K\top$

Monotonicity: $K(p \wedge q) \rightarrow (Kp \wedge Kq)$

Weak Necessitation: $Kp \rightarrow K\top$

Their main results are then:

Theorem 1 (DLR). *Let $M = \langle \Omega, k, a\rangle$ be a model on which Plausibility, KU-Introspection and AU-Introspection are valid.*

1(i) *If Necessitation is valid on M, then Ap is valid on M.*

1(ii) *If Monotonicity is valid on M, then $Kp \rightarrow Aq$ is valid on M.*

2 *If Weak Necessitation is valid on M, then $Kp \rightarrow Aq$ and $Ap \leftrightarrow Aq$ are valid on M.*

Our presentation of DLR's result differs in superficial respects from their original presentation. DLR do not present their constraints in terms of the validity of certain axioms. Thus, for example, instead of requiring KU-Introspection to be valid on M, they require $k - a(E) = \emptyset$ for all events E. However, it is a routine exercise to show that this condition is equivalent to the validity of our corresponding axiom. The same point holds for the other axioms. In short, our later models will not be escaping their triviality result by a sleight of hand which depends on this presentation.

One reason for the variant presentation is that it will facilitate the later exposition. It also serves to demonstrate that standard state space models as discussed here are equivalent to what are now commonly known as neighborhood or Scott-Montague frames (see [55] and [44]). It is well known that given certain restrictions on the function interpreting knowledge, this function can be turned into a binary relation among states along the lines of those used by [37] and [29]. This representation as a binary relation is, in turn, formally interchangeable with the "possibility correspondences" introduced by [2] (see also [3]) and used throughout economic theory (see, e.g. [11]).

2.3 Two Kinds of States

In response to their triviality results, DLR suggest distinguishing informally between "real" states and "subjective" states. As we understand it, this distinction can be explained as follows. An epistemic model makes predictions about how an agent or group of agents will or would behave in particular situations. The model makes predictions about these situations by including states which represent them. The real states in a model are the states which represent situations the model is intended to describe. The model predicts an agent will behave a certain way in a particular situation just in case the agent behaves that way in the real state which represents that situation. The predictions of a single model are given by what holds at all its real states; the behavioral theory of a class of models is given by what holds at all real states in all its models.

A state in an epistemic model is subjective if it figures in the specification of what the agent knows or is aware of at some real state. According to this way of understanding real and subjective states, states may be both real and subjective. Suppose we wish to represent an agent who knows that a particular coin will be flipped, but who will not learn the outcome of this coin flip. If our model is intended to make predictions no matter how the coin lands, the subjective states needed to specify the agent's knowledge (heads and ignorant, tails and ignorant) will be exactly the real states; every state will be both real and subjective. But as DLR recognize, there is no reason to require all states to play both roles. In the earlier example, if we only wanted to make predictions about the situation in which the coin comes up heads, we would not count one of the states (tails and ignorant) as a real state; even so, to represent the agent's ignorance given heads, a state where the coin comes up tails would still have to be included as a subjective state. The point can also be illustrated with a less artificial example. Consider an analyst who wishes to model the interactions of agents who are rational, but who do not believe each other to be rational. To represent the beliefs of these agents, the analyst must include subjective states in which the agents are irrational. But although she includes these subjective states, the analyst has no intention of eliminating the claim that the agents are rational from the predictions of her theory. Rather, it is understood informally that these subjective states are not real; they do not represent situations the analyst aims to describe.

Put in the terms of our presentation, DLR's proposal is to allow subjective states in which the law of excluded middle, $p \vee \neg p$, may not be true. DLR have no intention of eliminating the law of excluded middle from the predictions of their theory. Rather, they introduce these subjective states to specify the agent's knowledge and unawareness at real states where classical logic holds. The theory of DLR's models is given by what holds at these classical real states, not by what holds at all states whatsoever. Still, since classical tautologies may fail in DLR's subjective states, their models violate the "real states" assumption, and so are not standard state space models.

But once we acknowledge that some states may not be intended to represent situations the analyst wishes to describe, a question arises: Why should one require DLR's three axioms on unawareness to be valid in *all* states? DLR do not argue for this assumption. At best, their arguments in favor of their axioms only motivate imposing these axioms at real states. These arguments provide no motivation for imposing the axioms on subjective states that are not real, since these states are merely included in the model to specify the agent's knowledge and unawareness at real states.

There is a general methodological principle in epistemic modeling that *axioms* are to be imposed at all states. But in the literature on awareness, following DLR, this methodological principle has long been abandoned. DLR's own non-standard models violate this requirement, as do the current leading proposals for representing awareness, for example that of [24]. The subjective states in DLR's models include states in which logical axioms, including the law of excluded middle, do not hold. In DLR's models, as in the ones we will propose, even logical axioms are allowed to fail at some states. The only difference between our proposal and theirs concerns which axioms are allowed to fail. DLR preserve their own axioms at all states, and move to non-standard models in which classical propositional logic may fail at subjective states. We will preserve classical propositional logic at all states, and work with standard models in which DLR's axioms may fail at subjective states.

Still, one might ask: How should we *understand* a state where DLR's axioms are false? DLR interpret their own subjective states as "descriptions of possibilities as perceived by the agent" (p. 171). This interpretation does not seem appropriate for our models, in which DLR's axioms may fail at subjective states. But such metaphorical interpretations of these states are unnecessary. Subjective states where the axioms of awareness are invalid are simply to be understood in terms of the agent's knowledge and awareness at real states where the axioms are valid.

In fact, we can give a direct argument for not imposing DLR's axioms at all states, and in particular, for including states where KU-Introspection is invalid. First, by AU-Introspection, if an agent is unaware of p, then she must be unaware of being unaware of p. But then, by Plausibility, the agent does not know that she does not know that she is unaware of p. (Essentially, DLR already give this argument on p. 169.) In epistemic models, we generally represent an agent's not knowing q by including a state in which q is false. So, to allow for real states in which the agent does not know that she does not know that she is unaware of p, we must include subjective states in which the agent knows that she is unaware of p, and so violates KU-Introspection.[2]

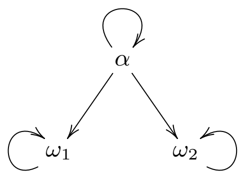
To sum up: after rejecting standard state space models, DLR propose that we should use models in which the laws of logic fail at subjective states. They implement this proposal by countenancing states where propositional logic fails, so that their models are non-standard. But if we allow models in which the law of excluded middle may fail at subjective states, we must also consider models in which other axioms, including DLR's, may fail at subjective states. DLR's formal results only apply to standard state space models in which their axioms are imposed at all states; the results do not concern standard state space models in which the axioms are imposed only at real states. As a consequence, these formal results cannot provide a basis for the conclusion DLR draw: that standard state space models preclude unawareness.

2.4 Non-Triviality

We have not argued against the validity of DLR's three axioms in real states – states representing the situations to be modeled. In our setting, we can formalize the distinction between real states and subjective states which are not real, by only assuming the validity of the axioms in some subset of the states in a model. We can then ask: is assuming the validity of the three axioms in such distinguished real states enough to lead to triviality? The following example shows that it is not:

THEOREM 2. *There is a model $M = \langle \Omega, k, a \rangle$, state $\alpha \in \Omega$ and event $E \subseteq \Omega$ such that Necessitation is valid on M, Plausibility, AU-Introspection and KU-Introspection are valid in α, and $\alpha \notin a(E)$.*

PROOF. Let $\Omega = \{\alpha, \omega_1, \omega_2\}$. Define a binary (accessibility) relation R on Ω as follows:



R induces a possibility correspondence P such that $P(\sigma) = \{\tau : R\sigma\tau\}$. With P, define k and a such that for all $F \subseteq \Omega$:

$$k(F) = \{\sigma \in \Omega : P(\sigma) \subseteq F\}$$

$$a(F) = \begin{cases} \{\omega_2\} & \text{if } \omega_1 \in F \text{ and } \omega_2 \notin F \\ \Omega & \text{otherwise} \end{cases}$$

It is routine to check that $M = \langle \Omega, k, a \rangle$, α and $E = \{\alpha, \omega_1\}$ witness the claim to be proven.[3] □

This shows that DLR's Theorem 1(i) cannot be extended to standard state space models in which DLR's three axioms are only required to be valid in real states. In fact, the model used in the above proof of Theorem 2 can also be used to

[2] This argument differs from DLR's main proof of triviality, since it only assumes that the axioms hold at the real state.

[3] As suggested in [43], one might consider an extension of Plausibility along the following lines: For each natural number n, let n-Plausibility be $Up \to (\neg K)^n p$, where $(\neg K)^n \varphi$ is defined inductively by the two clauses $(\neg K)^1 \varphi = \neg K \varphi$ and $(\neg K)^{n+1} \varphi = \neg K (\neg K)^n \varphi$. For each natural number n, n-Plausibility is valid in α.

show that DLR's other two results cannot be extended either. For 1(ii), note that Monotonicity is valid on M, and that $\alpha \in k(\Omega)$ although $\alpha \notin a(E)$. For 2, note first that Weak Necessitation is valid on M, and that $\alpha \in a(\Omega)$ but as before $\alpha \in k(\Omega)$ and $\alpha \notin a(E)$. More generally, any state in any model which satisfies both Necessitation and Monotonicity, in addition to DLR's three axioms, will be a counterexample not just to extensions of DLR's Theorem 1(i), but also to extensions of their Theorems 1(ii) and 2.

We conclude that none of DLR's three triviality results show that standard state space models preclude unawareness. One might wonder whether plausible strengthenings of the axioms on knowledge and unawareness allow us to reinstate the triviality results. In the full paper, we argue first that this cannot be achieved by strengthening their axioms governing knowledge, and, second, that it cannot be achieved by a particular strengthening of the axioms governing unawareness. The theorems and proofs may be found in Appendix A.

3. THREE KINDS OF AWARENESS

3.1 A New Triviality Result

DLR's result had limited implications for state space models because it depended on the validity of their axioms at all states. Is there a triviality result which only uses the validity of axioms on awareness in real states, rather than their validity in all states? In fact, as we now show, widely accepted axioms on awareness *do* lead to triviality even if they are imposed only at real states. The result uses the following two axioms:

AS: $A\neg p \to Ap$

AC: $A(p \wedge q) \to (Ap \wedge Aq)$

Awareness is widely assumed to satisfy both of these axioms; see, e.g., [43, pp. 274–275], [18, p. 331] (axioms A1 and A2) and [25, p. 309] (axioms 1 and 2).

As the next theorem shows, these axioms lead straightforwardly to triviality.

Theorem 3. *Let $M = \langle \Omega, k, a \rangle$ be a model and $\alpha \in \Omega$ such that AS and AC are valid in α. Then $Ap \to Aq$ is valid in α.*

Proof. Consider any events E and F, and assume $\alpha \in a(E)$. Since $E = E \cap \Omega$, $\alpha \in a(E \cap \Omega)$, and so by AC, $\alpha \in a(\Omega)$. By AS, $\alpha \in a(\emptyset)$. Since $\emptyset = \emptyset \cap F$, $\alpha \in a(\emptyset \cap F)$, and so by AC, $\alpha \in a(F)$. □

The crucial difference between this and DLR's triviality result is that AS and AC are only assumed to be valid in a distinguished state, for which it is shown that non-trivial unawareness in it is ruled out.

But does awareness really satisfy both AS and AC? In the following, we will focus in particular on AC, arguing that for some important notions of unawareness, being aware of a conjunction does not entail being aware of its conjuncts.

3.2 Attending vs. Conceiving vs. Processing

In the literature on awareness, it is uncontentious that there is no single attitude of awareness; what is expressed by "aware" is a loose cluster of notions. This was noted at the very start of the literature, as witnessed by the lengthy discussions in [10]; another detailed discussion can be found in [54]. We will argue that that at least some important notions of awareness do not satisfy AC (for others, as we will see, the situation is more complex).. In order to do so, we roughly distinguish the following three ways of understanding a claim of the form "The agent is aware of . . . ":

(i) The agent is attending to . . .

(ii) The agent has the conceptual resources required to conceive of . . .

(iii) The agent is able to process . . .

We will introduce these notions – and distinguish between them – using various examples found in the literature.

Consider first attention. An influential example, which first appeared in [16], and which is discussed at length by DLR and numerous places in the subsequent literature, is based on the following quote from one of Arthur Conan Doyle's Sherlock Holmes stories [9]:

> "'Is there any other point to which you would wish to draw my attention?'
> 'To the incident of the dog in the night-time.'
> 'The dog did nothing in the night-time.'
> 'That was the curious incident' remarked Sherlock Holmes."

Holmes's interlocutor is Inspector Gregory, a Scotland Yard detective. Before Holmes pointed out to Gregory that the dog did nothing in the night-time, Gregory was *unaware* of the dog doing nothing in the night-time. Gregory's state of unawareness is naturally understood as one of *inattention* – Holmes makes Gregory aware of the dog doing nothing in the night time in the sense of bringing this fact to his attention.

Gregory's failing to attend to the dog doing nothing in the night-time must be sharply distinguished from Gregory's not being able to conceive of the the dog doing nothing in the night-time. Before Holmes alerted Gregory to the dog doing nothing in the night-time, Gregory possessed the concepts required to entertain thoughts about the dog doing nothing in the night time. Contrast this with the following example for unawareness from [10, p. 40]:

> "How can someone say that he knows or doesn't know about p if p is a concept he is completely unaware of? One can imagine the puzzled frown of a Bantu tribesman's face when asked if he knows that personal computer prices are going down!"

The relevant state of unawareness in this example is not merely a matter of the agent failing to attend to the relevant event or subject matter. For example, if one is unaware in the sense of being unable to conceive of an event, it must be that one does not understand the words for those notions in any language. Contrast this with the case of Inspector Gregory. Gregory understands what Holmes says: he can conceive of the dog's doing nothing. But the purported example of inconceivability does not have this structure: the tribesman is supposed to be unable to think about computers using any of his conceptual resources, no matter what he attends to. The two notions of awareness – attending to versus being able to conceive of – are therefore clearly distinct.

The third notion of unawareness we want to single out is one which [10] (see also [19] and [20]) focus on; it can be understood as an attempt to deal with what is known as the "problem of logical omniscience" in epistemic logic. In standard state spaces, if two sentences φ and ψ are equivalent in classical propositional logic, then $K\varphi$ and $K\psi$ will be true in the same states. In particular, if $K(p \vee \neg p)$ is true in a given state, then so is $K\tau$ for any propositional tautology τ. One of Fagin and Halpern's reasons for developing a logic of awareness is to obtain logics which do not have this property. They write:

> "The notion of awareness we use in this approach is open to a number of interpretations. One of them is that an agent is aware of a formula if he can compute whether or not it is true in a given situation within a certain time or space bound. This interpretation of awareness gives us a way of capturing resource-bounded reasoning in our model."

Being unaware of φ in the sense of not being able to process φ is clearly distinct from failing to attend to φ: although Gregory did not attend to the dog doing nothing in the night-time, he had no difficulties processing the claim that the dog did nothing in the night-time. Not being able to process φ is also clearly distinct from not being able to conceive of φ: Gregory might not have been able to process an extremely complicated propositional tautology using only negation, conjunction and the sentence "the dog did nothing in the night-time", but he clearly possessed all the concepts required to entertain it.

3.3 Awareness of Conjunctions

Let's return to the new triviality result introduced at the start of this section. As already advertised, we believe that the principle AC, which says that an agent who is aware of a conjunction is aware of its conjuncts, may be plausible for one notion of awareness, but it is not for the other two.

Consider first awareness as the ability to process. This is plausibly a relation of agents to *sentences*, as part of what it takes to process ... is to be able to find out what the sentence "..." means. AC may hold if awareness is understood as the ability to process: It is natural to assume that an agent who is able to process a conjunction "... and —" is also able to process "..." and "—". As noted already in [10, p. 54], even this may fail: an agent might be able to recognize that a very long sentence has the form $\varphi \wedge \neg\varphi$, and so be able to process it, although she is unable to process the complex φ on its own. Resolving this controversy may require distinguishing further among different notions of processing, and the appropriate resolution may depend on the intended application.

The relations of attention and conceivability are different from the ability to process. In particular, they are plausibly relations agents have not to sentences but to what sentences express. We might call these entities *contents* or *propositions*, but in keeping with the terminology employed above, we call them *events*. In the full paper, we discuss the difference between sentences and events in more detail, and motivate the assumption of a *coarse-grained theory of events*, according to which events form a complete atomic Boolean algebra. A consequence of this assumption – which is implicit in all standard state-space based modeling techniques – is that sentences which are equivalent in propositional logic have identical contents.

With this understanding of attention and conceivability as relations of agents to coarse-grained events, consider first attention. Assume that after the conversation with Holmes quoted above, Gregory is alone thinking about the case, and attending to the event of the dog barking in the night (p). He is not, however, attending to event of Holmes at that moment smoking a pipe (q). It is then natural to say that Gregory is also not attending to the conditional event that *if* Holmes is currently smoking a pipe *then* the dog barked in the night ($q \to p$). But notice that according to the coarse-grained Boolean theory of events, the event that the dog barked in the night (p) is identical to the event that if Holmes is smoking a pipe then the dog barked in the night, and the dog barked in the night ($(q \to p) \wedge p$). So if AC were valid, then since Gregory is attending to the event of the dog barking in the night, he would be attending to the event that if Holmes is smoking a pipe then the dog barked in the night. But by assumption Gregory is not attending to this last event. Thus it follows from the coarse-grained theory of events that AC must be rejected.

A similar example can be given if we understand awareness as conceivability. Assume our agent does not have the conceptual resources to entertain the event of there being a black hole. According to the assumed coarse-grained theory of content, the event of there being a black hole and there being no black hole is identical to any event expressed by a propositional contradiction, such as the event of there being a sheep and there being no sheep. The agent might well have the conceptual resources to entertain the event of there being a sheep and there being no sheep, without having the conceptual resources to entertain the event of there being a black hole.

If we adopt, as usual, a theory of events which identifies the event expressed by sentences which are equivalent in propositional logic, AC appears to be inappropriate. Thus the new triviality result with which we started this section also does not establish that standard state space models preclude unawareness understood as inattention or inability to conceive.

4. PARTITIONAL MODELS

So far, we have shown that standard state-space models escape certain putative impossibility results for models of attention and conceivability. But this does not establish that standard space models can provide fertile models of these notions. In the remainder of the paper, we define, motivate, and examine a class of standard state space models for representing attention and the ability to conceive.

To show that our models generalize smoothly to the multi-agent case, from now on we use a language L_I parametrized to an arbitrary set of agent-indices I which is defined as the language L above, except that the operators A_i and K_i are indexed to $i \in I$. Models are consequently tuples of the form $\langle \Omega, k^i, a^i \rangle_{i \in I}$.

The models we will be working with are defined as follows:

Definition 1. *$\langle \Omega, R^i, \approx^i \rangle_{i \in I}$ is a* partitional model *if Ω is a set and for each $i \in I$, R^i is a binary relation on Ω which is reflexive and transitive, and $\approx^i$ a function which maps each $\omega \in \Omega$ to an equivalence relation $\approx^i_\omega$ on Ω.*

Here and in what follows, we make use of the fact that each equivalence relation corresponds to a unique partition, and *vice versa*; accordingly, we treat them as interchangeable.

Partitional models can be used to generate standard models in the following way: R^i specifies states of knowledge just as in Theorem 2. The idea behind $\approx^i$ is that the events the agent is aware of at ω are the events which are unions of sets of equivalence classes of $\approx^i_\omega$ (equivalently: unions of sets of cells of the induced partition). So for each $i \in I$, let R^i and $\approx^i$ determine functions k^i and a^i on events on Ω as follows:

$k^i(E) = \{\sigma \in \Omega : P^i(\sigma) \subseteq E\}$, where $P^i(\sigma) = \{\tau : R^i\sigma\tau\}$

$a^i(E) = \{\sigma \in \Omega :$ for all ρ and τ such that $\rho \approx^i_\sigma \tau$, $\rho \in E$ iff $\tau \in E\}$

Let the standard model determined by a partitional model $\langle \Omega, R^i, \approx^i \rangle_{i \in I}$ be $\langle \Omega, k^i, a^i \rangle_{i \in I}$, with k^i and a^i as just defined. On such a standard model, L_I can be interpreted as above; obviously, this induces a way of interpreting L_I directly on partitional models.

4.1 The Attitude of Attention

In order to motivate partitional models as models of limited attention, we suggest that attention in the sense we have been using the term should primarily be understood as an attitude towards questions. There are many available formal approaches to modeling questions (for an overview, see [35]). For concreteness, we'll adopt a standard approach, representing questions as partitions of the state space (see [17], building on [23] and [32]). Although we think the attitude to questions is primary, we will follow the literature on awareness, in axiomatizing a notion of attention which has events as its objects. The relationship between this attitude to events and the attitude toward questions will be as follows: an agent attends to the question Q if and only if the agent attends to every partial answer to Q. Using partitions to model questions, partial answers are unions of sets of cells, corresponding to how standard models are derived from partitional models.[4]

4.2 Partitions for Conceivability

To motivate the use of partitional models of conceivability, assume that the agents to be modeled have the concept of negation and (infinitary) conjunction, so that the set of events they can conceive of are closed under complement and arbitrary intersection. This is mathematically equivalent to requiring that this set is derived from an equivalence relation as above.

4.3 An Example

It will be useful to have a concrete partitional model before us, as a running example. The following model shows that there are non-trivial partitional models; for simplicity, a single-agent case is specified. Let $M = \langle \Omega, R, \approx \rangle$, with $\omega R \nu$ iff $\omega = 1$ or $\omega = \nu$, and $\approx$ given by the following equivalence classes:

$\approx_1$: $\{1\}, \{2,3,4\}$

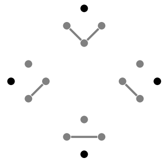
$\approx_2$: $\{1\}, \{2\}, \{3,4\}$

$\approx_3$: $\{1\}, \{3\}, \{2,4\}$

$\approx_4$: $\{1\}, \{4\}, \{2,3\}$

Thus, at 1, the strongest event known by the agent is Ω, and at each other state n, it is $\{n\}$. At each state n, the events the agent is aware of are the events which don't distinguish between any states in $\Omega \backslash \{1, n\}$.

Drawing the four states in a circle, starting with 1 at the top and going clockwise, we can draw each equivalence relation in a similar smaller circle, connecting two states by a sequence of lines if they are related by the relevant equivalence relation:



This is a partitional model in which there is non-trivial unawareness at each state. We will appeal to it below in order to show the consistency of various constraints.

4.4 Axioms

Given a class of models C, a set of sentences $\Sigma \subseteq L_I$ is the logic of C if and only if Σ contains exactly those sentences which are valid on C. Characterizing the logic of a class of models gives us a formal perspective from which to assess what assumptions our models encode about agents knowledge and awareness.

Thus we may ask: What is the logic of partitional models? Standard techniques on completeness results in modal logic are easily adapted to obtain the following result.

THEOREM 4. *A formula is valid on all partitional models if and only if it is derivable in the calculus with the following axiom schemas and rules:*

PL: Any substitution instance of a theorem of propositional logic.

K-K: $K_i(\varphi \to \psi) \to (K_i\varphi \to K_i\psi)$

K-T: $K_i\varphi \to \varphi$

K-4: $K_i\varphi \to K_iK_i\varphi$

A-Neg: $A_i\varphi \to A_i\neg\varphi$

A-M: $(A_i\varphi \wedge A_i\psi) \to A_i(\varphi \wedge \psi)$

A-N: $A_i\top$

K-RN: From $\vdash \varphi$ *infer* $\vdash K_i\varphi$

A-RE: From $\vdash \varphi \leftrightarrow \psi$ *infer* $\vdash A_i\varphi \leftrightarrow A_i\psi$

Moreover, the logic is decidable.

A proof is given in Appendix B.

4.5 DLR Once More

Consider again DLR's three axioms. Given our discussion above, it is natural to consider partitional models where DLR's axioms are required to be valid in some distinguished state:

[4] Section 5 discusses related models developed in [33, 34].

DEFINITION 2. *$\langle\Omega,\alpha,R^i,\approx^i\rangle_{i\in I}$ is a* partitional DLR model *if $\langle\Omega,R^i,\approx^i\rangle_{i\in I}$ is a partitional model and Plausibility, KU-Introspection and AU-Introspection (for each $i\in I$) are valid in α.*

We now show that DLR's triviality result cannot be revived in partitional models:

THEOREM 5. *There is a partitional DLR model $\langle\Omega,\alpha,R^i,\approx^i\rangle_{i\in I}$ and an event $E\subseteq\Omega$ such that $\alpha\in U(E)$.*

PROOF. Simply distinguish state 1 in the model presented in section 4.3. □

We conjecture that the logic of partitional DLR models can be axiomatized as follows:

CONJECTURE 1. *Add the following axioms to the theorems of the axiom system in Theorem 4 and close under modus ponens:*

P: $U_i\varphi\rightarrow(\neg K_i\varphi\wedge\neg K_i\neg K_i\varphi)$

AU: $U_i\varphi\rightarrow U_iU_i\varphi$

A formula is derivable in this calculus if and only if it is valid in every distinguished state of every partitional DLR model.

Note that $\neg KU\varphi$ can be derived using P, AU and K-T.

The present result shows that we can impose the DLR axioms without trivializing partitional models. But we confess to doubts about whether these axioms are appropriate. Just as with AC, once we understand more clearly the character of attention and conceivability, as well as the distinction between sentences and what they express, DLR's axioms become much less compelling. The clearest case concerns Plausibility and attention. A consequence of the contrapositive of Plausibility is $K\varphi\rightarrow A\varphi$. But this principle is false for attention. You know that there are more than four stars in the universe, but we doubt that you were attending to the question of how many stars there are prior to reading the previous clause. As we discuss in more detail in the full paper, the coarse-grained conception of content together with clarity about the notion of awareness to be modeled cast doubt on DLR's axioms.

4.6 Propositional Quantification

A challenge to some approaches to unawareness is to represent propositionally quantified statements. E.g., earlier models by Halpern made the claim that the agent knew she was unaware of something unsatisfiable (cf. [21] and [22]). In standard state space models such as ours, it is trivial to add propositional quantifiers without any such consequences. To do so, we write $v[E/p]$ for the valuation function which maps p to E and every other proposition letter q to $v(q)$:[5]

$$[\![\forall p\varphi]\!]_{M,v}=\bigcap\nolimits_{E\subseteq\Omega}[\![\varphi]\!]_{M,v[E/p]}$$

To illustrate that these quantifiers behave just as one would expect, note that in state 1 of the example described in section 4.3, the agent knows that she is unaware of something without there being something that she knows to be unaware of: $K\exists pUp\wedge\neg\exists pKUp$ is true in this state.

4.7 Closure and Automorphisms

In partitional models, what agents are aware of (attend to/can conceive) is closed under negation and conjunction. One might wonder whether we can also impose the constraints that what agents are aware of must be closed under awareness and knowledge. In other words, whether there are models on which the following axioms are valid:

A-4ij $A_ip\rightarrow A_iA_jp$

AK-4 $A_ip\rightarrow A_iK_jp$

To provide models which validate these principles we adapt the coherence constraint of [14].[6] The idea behind it is most easily described for awareness as conceivability, taking the equivalence relations of partitional models to represent a relation of indistinguishability using conceptual resources available to the relevant agent at the relevant state. Coherence requires that if two states are indistinguishable in this way, then there must be a way of permuting the state space in a way which preserves all structural facts about knowledge and awareness, as well as all the events which the relevant agent is aware of at the relevant state.

Let $M=\langle\Omega,R^i,\approx^i\rangle_{i\in I}$ be a partitional model. A permutation f of Ω is an automorphism of M if for all $i\in I$,

(i) for all $x,y,z\in\Omega$, $y\approx^i_x z$ iff $f(y)\approx^i_{f(x)} f(z)$, and

(ii) for all $x,y\in\Omega$, R^ixy iff $R^if(x)f(y)$.

A state $x\in\Omega$ *coheres* if for all $i\in I$ and $y,z\in\Omega$ such that $y\approx^i_x z$ there is an automorphism f of M such that $f(y)=z$ and $f\subseteq\approx^i_x$ (i.e., $\omega\approx^i_x\omega$ for all $\omega\in\Omega$). It's routine to verify that A-4ij and AK-4 are valid in any coherent state of a partitional model.

Once again, the model presented in section 4.3 demonstrates the satisfiability of this constraint: every state in this model is coherent. Since the model also satisfies the DLR axioms at state 1, it shows that even if we were to uphold the DLR axioms, imposing them together with coherence would not trivialize state space models of awareness.

5. DECISION THEORY

In section 4.5 the example of the number of stars illustrated how one may believe and know things to which one is not attending; clearly this kind of inattention may also affect choice-behavior. One advantage of standard state spaces is that we can use the usual decision-theoretic framework to represent the effects of inattention on choice-behavior.[7]

The usual decision theoretic representation of an agent's beliefs is given by a measure-space $\langle S,\mathcal{B},\mu\rangle$.[8] To generate a partitional model, we enrich this description of the agent by selecting $\mathcal{B}^C$, a complete atomic subalgebra of $\mathcal{B}$, to generate a representation of what the agent attends to in the context of choice: $\langle S,\mathcal{B},\mu,\mathcal{B}^C\rangle$. The atoms of $\mathcal{B}^C$ are a partition of S, so this structure gives rise to a partitional model of unawareness. The distribution the agent "uses" in a choice

[5] See already [36], and [12] for a more systematic development. See [13] for results on the complexity of propositional quantifiers in the related setting of [14].

[6] The following notion of coherence differs importantly from that of [14] in that $\approx^i_x$ here need not relate x only to x.

[7] We do not here attempt to back-form what the agent is aware of from her choice-dispositions, as [46, 47] do for belief, and [51, 52] do for awareness.

[8] This can be derived in any of the standard ways: e.g. [50], [59], [1], [5], [30], [6].

context is given by letting $\mu^C(E) = \mu(E)$ for all $E \in \mathcal{B}^C$ and undefined otherwise. The events the agent "explicitly believes" in the context can then be defined as the events of which the agent is certain in μ^C. The algebra $\mathcal{B}^C$ can also be used to parametrize "expanding" and more generally "changing" awareness, represented as transitions between different complete atomic sub-algebras of $\mathcal{B}$.[9] Since different algebras will determine different *explicit* beliefs in different contexts, this changing awareness can also represent effects of limited attention such as framing effects or failures of recall.

An approach along these lines has already proven fruitful in epistemic game theory. [33] and [34] develop Harsanyi type spaces in which players' beliefs may be defined on different σ-algebras. If the algebras are taken as the events the agent is attending to, one may interpret these models as examples of agents who fail to attend to questions about the higher-order beliefs of others, and thus do not have *explicit* beliefs over events which can be defined only by the level-n beliefs of others for large enough n.

5.1 Speculative Trade

An important test of approaches to unawareness has been how they fare with speculative trade ([26]). Building on the work of [2], [41] proved their famous "no-trade" theorem, illustrating the extreme strength of $S5$ knowledge together with a common prior. One aim of representing "bounded" agents such as those with limited attention is to escape such paradoxes (for this perspective, see [45], [38]). Accordingly, we now provide a partitional DLR model with a common prior in which speculative trade is possible.

As is well known (see [16], [49], [48]), the "no-trade" theorem does not hold in general if agents' accessibility relations R^i are merely transitive and reflexive, but are not required to form an equivalence relation. Plausibility is incompatible with the R^i forming an equivalence relation ([42]). Still, the DLR axioms together with partitional awareness models impose substantial further constraints, which might be thought to rule out speculative trade. We now construct a partitional DLR model to show that speculative trade can still occur in the presence of DLR's axioms and nontrivial unawareness.

Let the states be $W = \{1, 2, 3, 4, 5\}$ and the agents be Alice, A, and Bob B. The accessibility relations are defined so that: $1R^Ax$ iff $x \leqslant 3$; $5R^Ax$ iff $x \geqslant 3$ and otherwise wR^Ax iff $w = x$, while $R^B = W \times W$. The partitions of the agents are induced by $\approx_1^A = \approx_2^A = \{\{1\}, \{2, 3\}, \{4\}, \{5\}\}$; $\approx_4^A = \approx_5^A = \{\{1\}, \{2\}, \{3, 4\}, \{5\}\}$; and for all w, $\approx_w^B = \approx_3^A = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}$. The agents' common prior is the uniform one, and two acts f and g have utility as follows: $f(1) = f(5) = 1$, $f(2) = f(4) = 5$, $f(3) = 7$; $g(w) = 4$ for all $w \in W$. If the agents update by conditionalization on their implicit knowledge, then Alice will invariably maximize utility by choosing f (since in states $2, 3, 4$ she is certain it does better, and in states 1 and 5 she expects to gain $1/3 \cdot 1 + 1/3 \cdot 5 + 1/3 \cdot 7 > 4$). Bob meanwhile does not update at all, so that he strictly prefers g (since $4 > 2/5 \cdot 1 + 2/5 \cdot 5 + 1/5 \cdot 7$).

6. CONCLUSION

Standard state space models of attention and conceivability are at least as successful as current non-standard state space models. The non-standard models are, however, more complicated, and it is unclear that this complexity affords any advantages in predictive strength or accuracy. Standard state space models of these phenomena promise to lead to a rich and rewarding theory, posing technical and conceptual challenges, and offering connections to related work by linguists, philosophers and logicians – as well as work on bounded reasoning elsewhere in economic theory.

7. ACKNOWLEDGMENTS

Thanks to Eddie Dekel, Bart Lipman and Tim Williamson for very helpful discussions of drafts of this paper. Peter Fritz was supported by an AHRC/Scatcherd European Scholarship and the German Academic Exchange Service; Harvey Lederman acknowledges the support of the Clarendon Fund.

8. REFERENCES

[1] F. J. Anscombe and R. J. Aumann. A definition of subjective probability. *Annals of mathematical statistics*, pages 199–205, 1963.

[2] R. J. Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239, November 1976.

[3] R. J. Aumann. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28:263–300, 1999.

[4] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, 2001.

[5] E. D. Bolker. A simultaneous axiomatization of utility and subjective probability. *Philosophy of Science*, pages 333–340, 1967.

[6] J. Broome. Bolker-Jeffrey expected utility theory and axiomatic utilitarianism. *The Review of Economic Studies*, 57(3):477–502, 1990.

[7] B. F. Chellas. *Modal Logic: An Introduction.* Cambridge: Cambridge University Press, 1980.

[8] E. Dekel, B. L. Lipman, and A. Rustichini. Standard state-space models preclude unawareness. *Econometrica*, 66(1):159–73, 1998.

[9] A. C. Doyle. The adventure of Silver Blaze. *The Strand Magazine*, 1901.

[10] R. Fagin and J. Y. Halpern. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[11] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge*. MIT Press, 1995.

[12] K. Fine. Propositional quantifiers in modal logic. *Theoria*, 36:336–346, 1970.

[13] P. Fritz. Logics for propositional contingentism. unpublished.

[14] P. Fritz. Propositional contingentism. unpublished.

[15] S. Galanis. Unawareness of theorems. *Economic Theory*, 52:41–73, 2013.

[16] J. Geanakoplos. Game theory without partitions, and applications to speculation and consensus. Technical report, Cowles Foundation Discussion Paper, 1989.

[17] J. Groenendijk and M. Stokhof. *Studies on the Semantics of Questions and the Pragmatics of Answers*. PhD thesis, University of Amsterdam, 1984.

[18] J. Y. Halpern. Alternative semantics for unawareness. *Games and Economic Behavior*, 37(2):321–339, 2001.

[19] J. Y. Halpern, Y. Moses, and M. Y. Vardi. Algorithmic knowledge. In *Proceedings of the 5th*

[9] See the full paper for an alternative related to that of [31].

conference on Theoretical aspects of reasoning about knowledge, pages 255–266. Morgan Kaufmann Publishers Inc., 1994.

[20] J. Y. Halpern and R. Pucella. Dealing with logical omniscience: Expressiveness and pragmatics. *Artificial Intelligence*, 175(1):220–235, 2011.

[21] J. Y. Halpern and L. C. Rêgo. Reasoning about knowledge of unawareness. *Games and Economic Behavior*, 67(2):503–525, 2009.

[22] J. Y. Halpern and L. C. Rêgo. Reasoning about knowledge of unawareness revisited. *Mathematical Social Sciences*, 65(2):73–84, 2013.

[23] C. L. Hamblin. Questions in Montague English. *Foundations of Language*, 10:41–53, 1973.

[24] A. Heifetz, M. Meier, and B. C. Schipper. Interactive unawareness. *Journal of Economic Theory*, 130(1):78–94, 2006.

[25] A. Heifetz, M. Meier, and B. C. Schipper. A canonical model for interactive unawareness. *Games and Economic Behavior*, 62:304–324, 2008.

[26] A. Heifetz, M. Meier, and B. C. Schipper. Unawareness, beliefs, and speculative trade. *Games and Economic Behavior*, 77:100–121, 2013.

[27] S. Heinsalu. Equivalence of the information structure with unawareness to the logic of awareness. *Journal of Economic Theory*, 147(6):2453–2468, 2012.

[28] S. Heinsalu. Universal type structures with unawareness. *Games and Economic Behavior*, 83:255–266, 2014.

[29] J. Hintikka. *Knowledge and Belief.* Cornell University Press, 1962.

[30] R. C. Jeffrey. *The logic of decision.* Chicago: University of Chicago Press, second edition, 1983.

[31] E. Karni and M.-L. Vierø. "Reverse Bayesianism": A choice-based theory of growing awareness. *The American Economic Review*, 103(7):2790–2810, 2013.

[32] L. Karttunen. Syntax and semantics of questions. *Linguistics and Philosophy*, 1:3–44, 1977.

[33] W. Kets. Bounded reasoning and higher-order uncertainty. Submitted MS, September 2014.

[34] W. Kets. Finite depth of reasoning and equilibrium play in games with incomplete information. Submitted MS, February 2014.

[35] M. Krifka. Questions. In K. von Heusinger, C. Maienborn, and P. Portner, editors, *Semantics. An International Handbook of Natural Language Meaning*, volume 2, pages 1742–1785. Berlin: Mouton de Gruyter, 2011.

[36] S. A. Kripke. A completeness theorem in modal logic. *Journal of Symbolic Logic*, 24:1–14, 1959.

[37] S. A. Kripke. Semantical analysis of modal logic I: Normal modal propositional calculi. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 9:67–96, 1963.

[38] H. Lederman. People with common priors can agree to disagree. *The Review of Symbolic Logic*, pages 1–35, 2014.

[39] J. Li. Information structures with unawareness. *Journal of Economic Theory*, 144(3):977–993, 2009.

[40] M. Meier and B. C. Schipper. Bayesian games with unawareness and unawareness perfection. *Economic Theory*, 56:219–249, 2014.

[41] P. Milgrom and N. Stokey. Information, trade and common knowledge. *Journal of Economic Theory*, 26(1):17–27, 1982.

[42] S. Modica and A. Rustichini. Awareness and partitional information structures. *Theory and Decision*, 37(1):107–124, 1994.

[43] S. Modica and A. Rustichini. Unawareness and partitional information structures. *Games and Economic Behavior*, 27(2):265–298, 1999.

[44] R. Montague. Universal grammar. *Theoria*, 36:373–398, 1970.

[45] S. Morris. The common prior assumption in economic theory. *Economics and Philosophy*, 11(2):227–253, 1995.

[46] S. Morris. The logic of belief and belief change: A decision theoretic approach. *Journal of Economic Theory*, 69(1):1–23, 1996.

[47] S. Morris. Alternative definitions of knowledge. In M. O. L. Bacharach, editor, *Epistemic logic and the theory of games and decisions*, pages 217–233. Kluwer Academic Publishers, 1997.

[48] A. Rubinstein and A. Wolinsky. On the logic of "agreeing to disagree" type results. *Journal of Economic Theory*, 51(1):184–193, 1990.

[49] D. Samet. Ignoring ignorance and agreeing to disagree. *Journal of Economic Theory*, 52(1):190–207, 1990.

[50] L. J. Savage. *The Foundations of Statistics.* John Wiley and Sons, 1954.

[51] B. C. Schipper. Awareness-dependent subjective expected utility. *International Journal of Game Theory*, 42(3):725–753, 2013.

[52] B. C. Schipper. Preference-based unawareness. *Mathematical Social Sciences*, 70:34–41, 2014.

[53] B. C. Schipper. Unawareness—a gentle introduction to both the literature and the special issue. *Mathematical Social Sciences*, 70:1–9, 2014.

[54] B. C. Schipper. Awareness. In H. van Ditmarsch, J. Y. Halpern, W. van der Hoek, and B. Kooi, editors, *Handbook of Logics for Knowledge and Belief.* London: College Publications, forthcoming.

[55] D. Scott. Advice on modal logic. In K. Lambert, editor, *Philosophical Problems in Logic: Some Recent Developments*, pages 143–173. Dordrecht: D. Reidel, 1970.

[56] K. Segerberg. *An Essay in Classical Modal Logic.* Number 13 in Filosofiska Studier. Uppsala: Uppsala Universitet, 1971.

[57] G. Sillari. Models of awareness. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory (*LOFT *7).* Amsterdam: Amsterdam University Press, 2006.

[58] G. Sillari. Quantified logic of awareness and impossible possible worlds. *The Review of Symbolic Logic*, 1:514–529, 2008.

[59] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton: Princeton University Press, 1944.

A. MATERIAL SUPPORTING SECTION 2

A.1 Stronger Assumptions about Knowledge

The model used in Theorem 2 already validates a number of attractive axioms on knowledge, suggesting that strengthening DLR's constraints on knowledge is unlikely to yield an interesting triviality result. In particular, the following axioms are valid on the model:

Distribution: $(Kp \wedge Kq) \rightarrow K(p \wedge q)$

Anti-Necessitation: $\neg K \bot$

Reflexivity: $Kp \rightarrow p$

Positive Introspection: $Kp \rightarrow KKp$

We can show more systematically that any strengthening of the axioms of knowledge which rules out unawareness does so trivially in the same way as Negative Introspection does. On the very mild assumption that the agent doesn't know the contradiction $\bot$, we can characterize the conditions under which a given model for the knowledge of an agent can be extended to an unawareness model in which the agent is unaware of a given p at a given point α in which DLR's three axioms are valid. Let a model M' *extend* a *knowledge model* $M = \langle \Omega, k \rangle$ just in case $M' = \langle \Omega, k, a \rangle$ for some function $a : 2^\Omega \rightarrow 2^\Omega$.

Theorem 6. *Let $M = \langle \Omega, k \rangle$ be a knowledge model, $\alpha \in \Omega$ and $E \subseteq \Omega$ such that Anti-Necessitation is valid in α. M has an extension such that Plausibility, KU-Introspection and AU-Introspection are valid in α and $\alpha \notin a(E)$ if and only if*

(i) $\alpha \in -k(E) \cap -k - k(E)$, and

(ii) there is an event F such that $a \in F \cap -k(F) \cap -k - k(F)$.[10]

Proof. Assume first that (i) and (ii). Let $a : 2^\Omega \rightarrow 2^\Omega$ be defined so that $a(E) = a(F) = -F$, and $a(G) = \Omega$ for all other events G, and consider the model $M' = \langle \Omega, k, a \rangle$. It is routine to verify that Plausibility, KU-Introspection and AU-Introspection are valid in α, and $\alpha \notin a(E)$, appealing to Anti-Necessitation in the proof for KU-Introspection.

For the converse, note that (i) follows the validity of Plausibility in α. For (ii), let $F = -a(E)$. Then $\alpha \in F$, by AU-Introspection, $\alpha \notin a(F)$, and so by Plausibility, $\alpha \in -k(F) \cap -k - k(F)$. □

In particular, as long as the constraints on knowledge allow for there to be an event E and a state $\alpha \in E$ such that $\alpha \in -k(E) \cap -k - k(E)$, standard state space models and DLR's three axioms will not preclude non-trivial unawareness.

A.2 Stronger Assumptions about Awareness

These results demonstrate that no plausible strengthening of the axioms governing knowledge will re-instate triviality. But what if we strengthen the axioms on awareness themselves?

To investigate this issue formally, extend the language L by a unary operator CK for common knowledge. To define its interpretation on a model $M = \langle \Omega, k, a \rangle$, derive the following functions on events: $k^1(E) = k(E)$, $k^{n+1}(E) = kk^n(E)$, and $ck(E) = \bigcap_{1 \leq n < \omega} k^n(E)$.

$$[\![CK\varphi]\!]_{M,v} = ck([\![\varphi]\!]_{M,v})$$

With this, we consider the following additional axioms on awareness:

CK-Plausibility: $Ap \rightarrow CK(Up \rightarrow (\neg Kp \wedge \neg K \neg Kp))$

CK-KU-Introspection: $Ap \rightarrow CK(\neg KUp)$

CK-AU-Introspection: $Ap \rightarrow CK(Up \rightarrow UUp)$

These additional axioms are also compatible with non-trivial unawareness. In fact, they are valid in state α of the example in the proof of Theorem 2. More generally, Theorem 6 can be extended straightforwardly to these three additional axioms, given the weak assumption that Necessitation and Anti-Necessitation are valid:

Theorem 7. *Let $M = \langle \Omega, k \rangle$ be a knowledge model in which Necessitation and Anti-Necessitation are valid, $\alpha \in \Omega$ and $E \subseteq \Omega$. M has an extension such that Plausibility, KU-Introspection, AU-Introspection, CK-Plausibility, CK-KU-Introspection and CK-AU-Introspection are valid in α and $\alpha \notin a(E)$ if and only if*

(i) $\alpha \in -k(E) \cap -k - k(E)$, and

(ii) there is an event F such that $\alpha \in F \cap -k(F) \cap -k - k(F)$.[11]

Proof. We establish (i) and (ii) as in the proof of Theorem 6. Assuming (i) and (ii), we define a as in the proof of Theorem 6, where it is noted that Plausibility, KU-Introspection and AU-Introspection are valid in α, and $\alpha \notin a(E)$. For the CK-conditions, consider any event G such that $\alpha \in a(G)$. Then by construction of a, $a(G) = \Omega$. Therefore $a(G) \cup H = \Omega$ for any event H, and so $\alpha \in ck(a(G) \cup H)$ by Necessitation, which establishes the validity of CK-Plausibility and CK-AU-Introspection in α. For CK-KU-Introspection, note that by Anti-Necessitation, $k - a(G) = \emptyset$, so $-k - a(G) = \Omega$, from which $\alpha \in ck(-k - a(G))$ follows again by Necessitation. □

B. PROOF OF THEOREM 4

Proof. Since the formulas derivable in this calculus form a classical model logic in the sense of [56], we can apply the standard canonical model construction technique; in particular, consider the smallest canonical model (see [7, chapter 9], especially p. 254). Consider any formula φ not provable in the above calculus, and let Γ be the set of subformulas of φ closed under Boolean combinations. A standard filtration of the canonical model through Γ produces a finite model in which φ is false. It is routine to prove that the neighborhood function for A_i associates with each state a field of sets; since the model is finite this field is generated by an equivalence relation, as required. The above filtration can be chosen in such a way as to preserve the transitivity of the relation for K_i; reflexivity is preserved by any filtration (see, e.g. [7, chapter 3], especially p. 106, or [4] p. 80).

The above argument also establishes that the logic thus axiomatized has the finite model property and so is decidable. □

[10] This result also holds if we replace Plausibility by n-Plausibility, for all natural numbers n, and (i) and (ii) by the correspondingly iterated conditions.

[11] Again, we can extend this result to n-Plausibility for all n analogously to the extension in the previous footnote.

An Axiomatic Approach to Routing

Omer Lev
Hebrew University and
Microsoft Research, Israel
omerl@cs.huji.ac.il

Moshe Tennenholtz
Technion
moshet@ie.technion.ac.il

Aviv Zohar
Hebrew University and
Microsoft Research, Israel
avivz@cs.huji.ac.il

ABSTRACT

Information delivery in a network of agents is a key issue for large, complex systems that need to do so in a predictable, efficient manner. The delivery of information in such multi-agent systems is typically implemented through routing protocols that determine how information flows through the network. Different routing protocols exist each with its own benefits, but it is generally unclear which properties can be successfully combined within a given algorithm. We approach this problem from the axiomatic point of view, i.e., we try to establish what are the properties we would seek to see in such a system, and examine the different properties which uniquely define common routing algorithms used today.

We examine several desirable properties, such as robustness, which ensures adding nodes and edges does not change the routing in a radical, unpredictable ways; and properties that depend on the operating environment, such as an "economic model", where nodes choose their paths based on the cost they are charged to pass information to the next node. We proceed to fully characterize minimal spanning tree, shortest path, and weakest link routing algorithms, showing a tight set of axioms for each.

Categories and Subject Descriptors

C.2.1 [**Computer Systems Organization**]: Computer-Communication Networks—*Network Architecture and Design*; C.2.2 [**Computer Systems Organization**]: Computer-Communication Networks—*Network Protocols*; C.2.6 [**Computer Systems Organization**]: Computer-Communication Networks—*Internetworking*; G.2.2 [**Mathematics of Computing**]: Discrete Mathematics—*Graph Theory*

General Terms

Design, Algorithms, Theory

Keywords

Routing, Axiomatic Approach, Routing algorithms

1. INTRODUCTION

The proper way to distribute power, disseminate information, or establish hierarchies in organizations is an issue encountered whenever there is a large enough network of agents that needs to interact in an orderly manner. For example, when trying to establish efficient lines of communications between agents which all need to reach a central hub, there are various properties we may desire in our system. We might want the system to be able to handle small changes in connections without causing disruptions throughout the network; we may want it to be flexible when we change its parameters so that various routing options are possible, and more. Indeed, the search for the right communication structure has played a role in early work on the foundations of the area of multi-agent systems [7, 15, 5], based on classical work in organization theory [8, 16].

More concretely, examining networking, one of the most important aspects of the design of a communication network is the way it routes information through its physical links. Routing protocols, such as those used in packet switching networks, circuit switching, or ad-hoc networks are designed with many goals in mind. They must adapt to changing network conditions, withstand failures, and operate in a distributed fashion while constructing a "good" routing scheme. Nodes in the network are, in fact, autonomous agents that can control the flow of information through them and can choose to forward it according to their own considerations. Agents may be controlled by different economic entities (such as in the internet, where different internet service providers control some of the routers), and may route according to complex preferences that are derived from economic relations [9, 14]. Even in the cooperative local-network setting where all routers are controlled by a single network operator, different considerations such as bandwidth utilization, latency, and the risks of link failures come into play.

The multitude of previous treatments of the problem suggest a myriad of routing protocols, each with their own benefits and shortcomings. In contrast, this work examines the routing problem through the lens of the *axiomatic approach*, which seeks to formulate different elementary properties that are desirable in this context. One approach to an axiomatic treatment, which we take in this work, is that of characterization: a set of elementary properties is shown to uniquely determine some routing algorithm, and hence the routing outcome on any specific graph. From the designer's perspective, such a result implies a great deal – any additional property that is not already achieved by the protocol cannot be added to it without giving up on another basic property. The approach thus provably bounds the design space of algorithms and makes explicit the choices made when selecting one over the other.

As we are not aware of any previous axiomatic treatment of routing, we focus our attention on a domain that most closely resembles the internet as it is built today, and focus our efforts within this domain on what one may consider classic, or natural routing schemes. In particular, we assume that routing choices are independent of the congestion on links (such is the case in the internet, where routing protocols such as BGP first establish paths, and congestion control protocols such as the one embedded into TCP manage the load on each flow's path and ensures that rates are throttled to match the bottleneck of the flow). Furthermore, as with internet routing where routers decide on the next hop of each packet using a routing table that maps its destination to the next hop, routing choices made to different destinations are done independently. Finally, packets addressed to the same destination are not split between different paths, and are routed in the same manner regardless of their source. These choices, which greatly restrict the power of any routing algorithm may seem arbitrary, but are in fact derived from real-world design considerations. For example, the need to quickly forward packets towards their destination at each router mandated that most routing be done in specialized hardware. No complex computation is performed (only a lookup into a routing table) and no deep inspection of the packet is performed. Keeping routing simple has made it fast and robust.

More advanced routing schemes that have been proposed in the literature may split traffic, allow routing choices to depend on the source of the packet or its previous hops, or may even change the routes in response to link congestion. These are notoriously difficult to coordinate and to implement. We leave treatment of these more advanced schemes to future work.

Our set of axioms or "desirable properties" are also motivated by similar considerations. For example, one of the fundamental features we desire in our algorithms is one of **robustness**, which is the ability of a system to endure changes in the network without creating disruption in parts of the network that have not undergone changes.

A different feature, which might be desirable only in certain cases, is **"first hop"**, which is particularly relevant for diffuse networks with independent nodes. It means, broadly, that network nodes care only about their immediate surroundings, or the "next step" in the network data transfer. Such a property might be relevant when nodes pursue an "economic model", paying for transferring information, and hence only caring about the cost they need to pay to move their information to the next node, and following that, they have no preference on the route the information should pass en route to its destination. Other properties, desirable only in some cases include an indifference between two parallel paths, as long as they change their weights by the same amount concurrently.

Ultimately, after devising our axioms we successfully fully characterized 3 natural routing algorithms:

- **Minimum spanning tree:** A tree with the smallest overall weight is a result, among others, of the "first hop" axiom (the "economic model").
- **Shortest path:** A tree where each node has the shortest possible path to its destination is a result, among other axioms, of viewing as immaterial to the routing decision any parallel paths which change their weight by the same amount.
- **Weakest link tree:** A tree where each node takes the path with the maximal "lightest" weight available to it. This results from considering higher edge-weights as beneficial (e.g., representing bandwidth which one wishes to increase in contrast to delay that one wishes to decrease), and from considering designers that choose between parallel paths in a slightly different manner.

We proceed to review relevant previous research and then continue to define our model and expand on the axioms, which are motivated with a brief explanation and presented formally. Following that we show (and prove) our characterization of the minimal spanning tree, the shortest path tree, and the weakest link tree.

2. RELATED WORK

In the past decade, as routers became more flexible, research on routing (particularly inter-domain) and its techniques has been rekindled and extended beyond the technical issues dealt with in the past. The harbinger for much of this research was [10], which was further expanded by several researchers (see updating report here: `http://www.cl.cam.ac.uk/~tgg22/metarouting/`). However, this line of research, while introducing many interesting mathematical and theoretical concepts to the field of routing, has refrained from phrasing its models as requirements by users, to be filled by various routing algorithms.

The axiomatic approach, which does approach problems with this outlook, has been first introduced in CS contexts as extensions to the classical theory of choice [4], and has been applied to ranking systems [1, 2] and trust systems [3], as well as to other multi-agent setups such as multi-level marketing [6].

In relation to networking, usage of the axiomatic approach has generally been concentrated in two main areas: applying to general graph theory (e.g., [18]) or in more technical approaches to networks: papers such as [13] which deal with particular wireless models and implementations, and, somewhat closer to our line of work, [12], whose basic axioms are basic enough to be covered through our models, while the routing related axioms involve various assumptions on how routers work (tables, etc.), which we refrain from approaching in our more abstract considerations.

Further work connecting networking and the axiomatic approach has focused on particular instances of problems: [11] try to use the axiomatic approach to extract the costs of multicast routing and decide who is to pay them. Trust networks and social networks (e.g., recommendation systems) have been analysed many times using the axiomatic approach to understand their desirable features and better understand desirable algorithms in these cases [17, 3]. However, none of these papers deal with the basic routing mechanism by which messages and information arrive at each node.

3. SETUP

Before introducing our axioms, we begin by setting up our routing model. It is, naturally, only a simplification of routing as it is done in large, complex networks such as the internet, but we believe it is robust enough to display many networking characteristics.

Our world will be a weighted graph $G(V, E, W)$ and a destination d, where V is a set of nodes, E is a set of edges, and W is a function assigning weights to edges, and $d \in V$. A routing solution is a tree T over that graph, as defined below (we do not concern ourselves with non-tree routing, as passing through the same node several times does not serve any purpose).

Definition 1. *A routing function $f_d : \mathcal{G} \to \mathcal{T}$ is a function from connected weighted graph $G(V, E, W) \in \mathcal{G}$ in which $d \in V$, to a tree $T(V, E, W) \in \mathcal{T}$ such that $T \subseteq G$.*

We can look at the graph as one with directed edges if we consider each edge's direction to be the one pointing at the vertex from which there is a path to d (without going through the same edge again).

We discuss 3 different routing options:

- *Minumum spanning tree (MST)*: a tree connecting all nodes in the graph with the minimal weight, i.e., for every tree $T' \subseteq G$ that encompasses all of G's nodes, $\sum_{e \in f_d(G)} W(e) \leq \sum_{e \in T'} W(e)$.
- *Shortest path*: each node is connected to d using a shortest length path in the graph. For every node $v \in V$, let $(e_1, \ldots, e_s)$ be a path without cycles from v to d such that $e_i \in T$, and let $(e'_1, \ldots, e'_k)$ a different path from v to d, then $\sum_{i=1}^{s} W(e_i) \leq \sum_{j=1}^{k} W(e'_j)$.
- *Weakest link*: looking at each potential path from each node to d, we give each path the value of its smallest valued edge. The routing tree will contain, for each node the path to d with the maximal value. So for every node $v \in V$, let $(e_1, \ldots, e_s)$ be a path without cycles from v to d such that $e_i \in T$, and let $(e'_1, \ldots, e'_k)$ a different path from v to d, then $\min_{1 \leq i \leq s} W(e_i) \geq \min_{1 \leq j \leq k} W(e'_j)$.

Notice that while for the minimal spanning tree and shortest path routing options weights are interpreted as costs (e.g. payments, delays), so these algorithms seek to minimize them, the weakest link views weights as measure for capability such as bandwidth, so seeks to maximize the weight.

4. AXIOMS

Having introduced our framework, we introduce our axioms, which are, basically, desirable properties of the function f_d (in the axioms below we use f, as these are properties which do not depend on a specific d destination).

Robustness indicates the routing being quite unsusceptible to changes – only if a path in the routing is destroyed, will it require any change. As indicated in Figure 1, the path from node a changes, but not from node b.

Axiom 1 (Robustness). *f is* robust *if removing an edge $e \in E$ from $G(V, E, W)$, yielding G', then for every vertex $v \in V$: if the cycle-less path from v to d in $f_d(G)$ did not contain e, then this is still the selected path according to $f_d(G')$ (see example Figure 1).*

The following axioms deal with global changes to the graph weights, additive or multiplicative:

Axiom 2 (Scale Invariance). *f is* scale invariant *if for a graph $G(V, E, W)$, for any positive scalar $\alpha \in \mathbb{R}_+$, defining $G'(V, E, \alpha W)$, for every $d \in V$, $f_d(G) = f_d(G')$.*

Axiom 3 (Shift Invariance). *f is* shift-invariant *if for a graph $G(V, E, W)$, for any $\alpha \in \mathbb{R}$, defining $G'(V, E, \alpha + W)$, for every $d \in V$, $f_d(G) = f_d(G')$.*

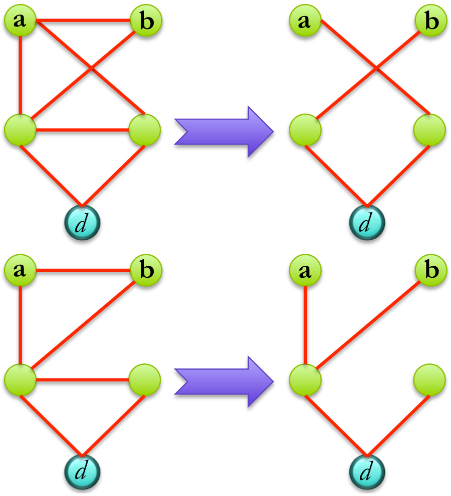


Figure 1: An edge is removed, but only a, whose path used that edge changes its path (the left side is the graph, the right side is the routing algorithm's output)

The monotonicity axiom below seeks to establish that if an edge does not have to be in every tree, if its weight increases enough, it will not be a part of the routing tree:

Axiom 4 (Monotonicity). *f is* monotone *if for a graph $G(V, E, W)$ and $d \in V$, for $e' \in E$, if $e' \notin f_d(G)$, then for every $G'(V, E, W')$, there is a value $M_{W'}$ such that for W'' such that $W''(e) = W'(e)$ for all $e \in E \setminus \{e'\}$ and $W''(e') \geq M_{W'}$, $e' \notin f_d(G''(V, E, W''))$. Similarly, we can define the opposite direction, an edge in $f_d(G)$ will not be in the routing tree if it has a small enough value; we will refer to it as* inverse monotonicity.

While the phrasing of the following axiom is somewhat technical, the *first hop* axiom below simply means that if a vertex has several potential edges to connect to a path to d, the routing only depends on the weights of the edges connecting it to these potential paths, and unrelated to weights of other edges in the graph.

Axiom 5 (First Hop). *Let $G(V, E, W)$ be a weighted graph and let $v, d \in V$ and $d \neq v$. Suppose $C = \{c_1, \ldots, c_s\}$ are the vertices such that $(v, c_i) \in E$ and there is a path from c_i to d in $f_d(G)$ which does not pass through v. W.l.o.g., let (v, c_1) be the first step in the path from v to d in $f_d(G)$. We say that f satisfies* first hop *if for any W' such that $W'(v, c_i) = W(v, c_i)$ and if for all $c_i \in C$ $f_d(G'(V, E, W'))$ contains paths to d from c_i that do not pass through v, and there is no $c' \notin C$ such that $(v, c') \in E$ and there is a path from c' to d in $f_d(G')$, then the cycle-less path from v to d in $f_d(G'(V, E, W'))$ starts with (v, c_1).*

The rational for the *first hop* axiom is to capture a common economic model, in which edge weights indicate the cost of passing information. In distributed networks, such as the internet, each agent only minds the amount it needs to pay to transfer its data to the next node, not caring about the path the data will take from there.

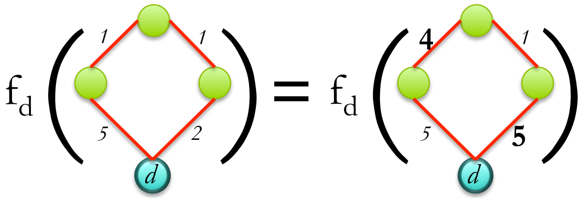


Figure 2: Selected path does not change when each path from the top node is added 2.

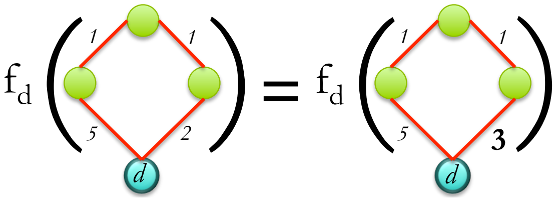


Figure 3: Selected path does not change when the bottom right edge is slightly increased.

Path cardinal/ordinal invariance intends to see the planner's considerations when multiple paths exist. As there might be many potential behaviours, we only limit ourselves to examining the narrow case of what the planner considers important when there is only one cycle in the graph (i.e., the axiom does not strongly enforce a general behaviour on the planner). Cardinal invariance deals with adding the same weight to potential paths, and how it does not effect the routing. Ordinal invariance similarly does not change the routing if all that has changed are the weights of the competing paths, as long as edges in each path maintain their relative position.

Axiom 6 (Path Cardinal Invariance). *Let $G(V, E, W)$ be a graph which contains a single cycle, $d \in V$, and let $d \neq v \in V$ be a part of this cycle. Hence there are two alternative paths from v to d – $p_1 \subset E$ and $p_2 \subset E$ (one of them is actually a part of $f_d(G)$). f is* path cardinal invariant *if it treats those paths as such: Choosing an edge $e' \in p_1$ and $e'' \in p_2$, for any $\alpha \in \mathbb{R}$, we define W' as $W(e) = W'(e)$ for $e \in E \setminus \{e' \cup e''\}$ and $W'(e') = W(e') + \alpha$ and $W'(e'') = W(e'') + \alpha$, the path from v to d will not change in $f_d(G(V, E, W'))$ (see example Figure 2).*

Axiom 7 (Path Ordinal Invariance). *Let $G(V, E, W)$ be a graph which contains a single cycle, $d \in V$, and let $d \neq v \in V$ be a part of this cycle. Hence there are two alternative paths from v to d – $p_1 \subset E$ and $p_2 \subset E$ (one of them is actually a part of $f_d(G)$). f is* path ordinal invariant *if it treats those paths as such: Taking an edge $e' \in p_i$ ($i \in \{1, 2\}$) that is not maximal or minimal in $p_1 \cup p_2$, we define W' as $W(e) = W'(e)$ for $e \in E \setminus \{e'\}$ and allow W'(e') to be any value it chooses as long as for every $e'' \in p_i$ if $W(e') \geq W(e'')$ then $W'(e') \geq W'(e'') = W(e'')$, and the path from v to d will not change in $f_d(G(V, E, W'))$ (see example Figure 3).*

5. MINIMAL SPANNING TREE

Theorem 1. *A robust, scale invariant, shift invariant, monotone, first-hop (axioms 1-5) routing function f, for any graph $G(V, E, W)$ and $d \in V$, $f_d(G)$ will always be a minimal spanning tree of G.*

Reminder 1. *As our minimal spanning tree proof relies on the Kruskal algorithm, we will briefly describe it:*

1. *Order edges according to weights*
2. *Define a set S, initialized to the empty set.*
3. *Going over edges from lightest to heaviest, if the set $S \cup \{e\}$ has no cycles, $S = S \cup \{e\}$.*

Proof Proof of Theorem 1. We shall prove the theorem using complete induction on the number of non-cycle lightest edges in the tree $f_d(G)$. Hence, we shall begin by proving that the lightest edge in the graph G is in the routing tree $T = f_d(G)$. Assuming we are mistaken, let us consider the lightest edge in G – $e = (u, v) \in E$ – and assume $e \notin T$. We create $G'(V, E', W) = f_d(G) \cup \{e\}$, and thanks to the robustness axiom, we know $f_d(G) = f_d(G')$.

If v's path to d in $f_d(G')$ goes through u, we shall switch the nodes' names, so that v's path to d does not pass through u. As e is not in $f_d(G')$, there is an edge $e' = (u, s)$ that is the first step from u towards d. We now define $x = W(e)$ and $y = W(e')$, and due to our minimality assumption, we know $x < y$.

Using the monotonicity axiom, we change graph weights to W' that is identical to W except that e' weight is large enough so that we create a tree T' in which there is a path from s to d that does not pass through u (e.g., the same path that is in $f_d(G')$), and v passes through u towards d (i.e., $e \in T'$). We define $y' = W''(e')$.

Using scale invariance we now multiply all edges by $\frac{y-x}{y'-x}$, and using shift invariance, we add to all edges $y - \frac{y-x}{y'-x}y'$. This means the weight of edge e is now

$$x\frac{y-x}{y'-x} + y - \frac{y-x}{y'-x}y' = (x - y')\frac{y-x}{y'-x'} + y = x$$

While the weight of edge e' is now

$$y'\frac{y-x}{y'-x} + y - \frac{y-x}{y'-x}y' = y$$

However, the routing tree contains e and not e', and a path from both v and s to d, contradicting the "first hop" axiom, which should have caused e' to be chosen over e, as the edge weights for e and e' have not changed.

We now turn to the induction step – we assume all bottom weighted $k-1$ edges that do not create a cycle are included in the tree $T = f_d(G)$, and we now seek to include the k-lightest edge that does not create a cycle. We pursue a similar path as we did as previously, and we shall mark the edge as $e = (u, v)$, and assume it is not included in $T = f_d(G)$, and instead $e' = (u, s)$ is included, and there is a path to d from v and s. Again, we create $G'(V, E', W) = f_d(G) \cup \{e\}$, and thanks to the robustness axiom, we know $f_d(G) = f_d(G')$. Using monotonicity we create weights W' that just increase e' weight, so that $G'' = (V, E', W')$ has $T' = f_d(G'')$ which include the same bottom k which do not create cycles (from the induction hypothesis), and u reaches d via the edge e. Recall that we know the bottom $k-1$ edges will definitely be in $f_d(G'')$, and we wish to ensure that there will still be a path from v to d and from s to d. The same arguments used in the initial step of the induction ensure that, as well as returning the weights of e and e' to their values in G, while routing u through e and not e' in the routing tree, reaching a contradiction with our initial assumption due to the "first hop" axiom.

What is left is to show MST indeed follows our axioms:

Robustness (axiom 1) Trivial thanks to the Kruskal algorithm – if the removed edge (e') was not in the routing tree, it means it was not selected in the first place, and hence the same routing tree will be chosen. If it was, then any edge added after its removal (e'') closed a cycle with it, and hence, if affecting the edges in any path that did not include e', it means e'' closes a cycle with them, hence e' would have closed a cycle as well.

Scale invariance (axiom 2) Multiplying all edges by a fixed amount does not change their order in relation to others, hence Kruskal will choose the same routing tree.

Shift invariance (axiom 3) Adding a fixed amount to all edges does not change their order in relation to others, hence Kruskal will choose the same routing tree.

Monotonicity (axiom 4) Giving an edge the maximal possible edge value ensures it will only be selected if no other edge can replace it – and if there exists a tree without some edge, we know it will be chosen before.

"First hop" (axiom 5) Kruskal ensures that if there are the same possible options of connecting a node to the tree, only the lightest edge will be chosen.

□

6. SHORTEST PATH

Theorem 2. *A robust, scale invariant, monotone, and path cardinal invariant (axioms 1-2, 4, 6) routing function f, for any graph $G(V,E,W)$ and $d \in V$, $f_d(G)$ will always be a shortest path graph to d of G.*

Proof. Suppose $T = f_d(G)$ is not a shortest path routing tree. Let u be the closest node to d that is not connected to d with a shortest path. Hence, there is an edge $e = (u,v)$ which will make u's path a shortest path one (v, being closer to d, is already connected to d with a shortest path), but $e \notin T$, and instead $e' = (u,s)$ is included in T. Using robustness, we create $G'(V,E',W) = T \cup e$. G' contains two alternate paths from u to d, and $f_d(G) = f_d(G')$.

Using path cardinal invariant, we "move" all the value of the edges on each path to it's "source", i.e., to (u,v) or (u,s) (we do this by adding to the weight of (u,v) and (u,s) the value of $\sum_{e\in(p_1\cup p_2)\setminus(p_1\cap p_2)} W(e) - W((u,v)) - W((u,s))$, and reduce from $W((u,s))$ the weight of all edges of the path from u to d through (u,v) and vice versa). We shall refer to $W(e) = x$ and $W(e') = y$. We now use monotonicity to create a new tree, with e but without e', with the graph's weight now W' (identical to W except for increase in e' weight). Once again, we transfer all value of the paths from u to d to e and e' respectively, with everything else being 0. Now, using monotonicity, we increase the weight of e' above that of e, with the weight of (u,v) being x (its path weights have not changed) and (u,s) being y'.

Finally, we multiply all edges by $\frac{y-x}{y'-x}$ (using scale invariance), and using path cardinal invariance, we add to e and e' the amount $y - \frac{y-x}{y'-x}y'$. The weight of e is now:

$$x\frac{y-x}{y'-x} + y - \frac{y-x}{y'-x}y' = (x-y')\frac{y-x}{y'-x} + y = x$$

While the weight of edge e' is now

$$y'\frac{y-x}{y'-x} + y - \frac{y-x}{y'-x}y' = y$$

As all edges are the same weight as before, therefore we reached a contradiction regarding the inclusion of e' instead of e (whose weights are the same as well).

We will now show shortest path follows our axioms:

Robustness (axiom 1) Removing an edge, at most, eliminates a potential path from a node to the destination d. If the path was not on the shortest path, the previous shortest path remains so.

Scale invariance (axiom 2) Multiplying by a fixed amount all edges means the value of each path is multiplied by the same amount, maintaining their relative ordering, hence what was shortest remains so.

Monotonicity (axiom 4) Giving an edge the value of the sum of all other edges ensures it will only be selected if no other path can replace it — and if there exists a tree without some edge, we know there is such a path.

Path cardinal invariance (axiom 6) Having multiple paths from a node, adding the same amount to each path doesn't change the ordering of the paths (i.e., which path is "shorter" than another), hence selection of shortest path will be identical.

□

7. WEAKEST LINK

Theorem 3. *A robust, scale invariant, shift invariant, inverse monotone, and path ordinal invariant (axioms 1-4, 7) routing function f, for any graph $G(V,E,W)$ and $d \in V$, $f_d(G)$ will always be a weakest link graph to d of G.*

Proof. Suppose $T = f_d(G)$ is not a weakest link routing tree. Let u be a node that requires just one edge missing from T that is not connected to d with a weakest link[1], and we mark this edge as $e = (u,v)$. Since $e \notin T$, there is an edge instead $e' = (u,s)$ that is included in T. Using robustness, we create $G'(V,E',W) = T \cup e$. G' contains two alternate paths from u to d, and $f_d(G) = f_d(G')$.

Using path ordinal invariant, we change the value of all edges on each alternate path from u to d to its weakest link value (we do this by taking the 2nd smallest edge in the path and changing its value to that of the weakest link, which by

[1] such a node exists as there is a node not connected by weakest link in T, hence adding the necessary path for that node, taking the node just before the final edge that we add to T (i.e., closest to d), answers our criterion.

the axiom does not change the path chosen, and we proceed doing so to all edges on the path). We shall refer to $W(e) = x$ and $W(e') = y$ (from assuming u is not in a weakest link path we know $x > y$). Using inverse monotonicity, we create W' identical to W except for e' weight, that is low enough that it is not included in $f_d(G''(V, E', W'))$. Once again, we change the values of the paths from u to d to their weakest link value (this is only relevant for the path through e', as the other path has not changed). We term the the new value for $e' - y'$, and we know $x > y'$.

Finally, we multiply all edges by $\frac{y-x}{y'-x}$ (using scale invariance), and using shift invariance, we add to all edges $y - \frac{y-x}{y'-x}y'$. Edge e now has the weight:

$$x\frac{y-x}{y'-x} + y - \frac{y-x}{y'-x}y' = (x - y')\frac{y-x}{y'-x} + y = x$$

While the weight of edge e' is now

$$y'\frac{y-x}{y'-x} + y - \frac{y-x}{y'-x}y' = y$$

As all edges are the same weight as before, hence we reached a contradiction regarding the inclusion of e' instead of e.

We shall now show weakest link also follows our axioms:

Robustness (axiom 1) Removing an edge, at most, eliminates a potential path from a node to the destination d. If the path was not a weakest link, the previous weakest link remains so.

Scale invariance (axiom 2) Multiplying by a fixed amount all edges means the value of each path (its smallest edge) is multiplied by the same amount, maintaining their relative ordering, hence what was weakest link remains so.

Shift invariance (axiom 3) Adding a fixed amount all edges means the value of each path (its smallest edge) is added the same amount, maintaining their relative ordering, hence what was weakest link remains so.

Monotonicity (axiom 4) Giving an edge the value of the minimum of all other edges ensures it will only be selected if no other path can replace it — and if there exists a tree without some edge, we know there is such a path.

Path ordinal invariance (axiom 7) Having multiple paths from a node, the weakest link edge (the one with smallest value) of the selected path can't become lower than the weakest link of the non-selected path, hence weakest link choice does not change.

□

8. TIGHTNESS OF AXIOMS

We will now show that the above characterizations are tight, and that without each axiom, other routing algorithms become possible.

THEOREM 4. *All MST axioms (1-5) are necessary, and without even one of them, other routing algorithms are possible.*

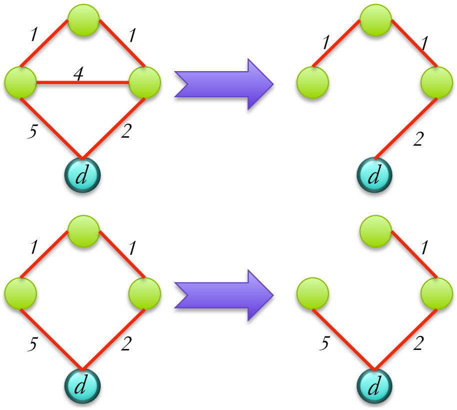


Figure 4: Lack of robustness results in a minimal spanning tree/shortest path routing (above) ending up in a routing tree that is weakest link but not MST or shortest path (below).

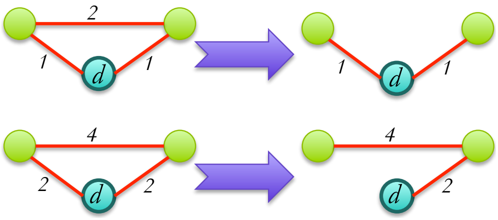


Figure 5: Eliminating scale invariance results in a minimal spanning tree/shortest path/weakest link routing (above) ending up in neither (below).

PROOF. Going over all MST axioms, we detail potential algorithms which work with all axioms except that one, and are not MST. We will refer below to each relaxed axiom, and to the new/additional system which can obtained by that relaxation:

Robustness See example in Figure 4. Apply MST to any other graph that is not a linear transformation of the bottom one.

Scale invariance See example in Figure 5. On all graphs except those which contain as a subgraph a linear transformations of the bottom one, apply MST.

Shift invariance See example in Figure 6. On all graphs except those which contain as a subgraph a linear transformations of the bottom one, apply MST.

Monotonicity A maximal spanning tree implements all axioms but monotonicity.

First Hop Weakest link implements all of the other axioms.

□

THEOREM 5. *All shortest path axioms (1-2, 4, 6) are necessary, and without even one of them, other routing algorithms are possible.*

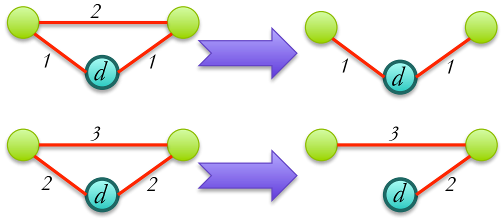


Figure 6: Eliminating shift invariance results in a minimal spanning tree/weakest link routing (above) ending up in neither (below).

PROOF. Going over all shortest path axioms, we detail potential algorithms which work with all axioms except one, and are not shortest path. We will refer below to the each relaxed axiom, and to the new/additional system which can obtained by that relaxation:

Robustness See example in Figure 4. Apply shortest path to any other graph that isn't a scale of the structure of the bottom one. Any edge in that structure that is 100 times all the others is removed in the tree.

Scale invariance See example in Figure 5. Taking the bottom example and for the group that includes all graphs for which it is a subgraph and those that can be formed by path cardinal invariance, and only for them do not apply shortest path but rather the example (it won't trample on the top example, as if the upper example adds y to lower-right edge, and y to the rest, and the bottom example adds x, it would require $2+x = 4+x$, reaching an impossibility).

Inverse Monotonicity A longest path tree implements all axioms but monotonicity.

Path cardinal invariant Minimal spanning tree implements all other axioms.

□

THEOREM 6. *All weakest link axioms (1-4, 7) are necessary, and without even one of them, other routing algorithms are possible.*

PROOF. Going over all weakest link axioms, we detail potential algorithms which work with all axioms except one, and are not weakest link. We will refer below to the each relaxed axiom, and to the new/additional system which can obtained by that relaxation:

Robustness See example in Figure 4. Apply weakest link to any other graph that isn't of the structure of the bottom one, Any edge that in that structure that is 100 times less that all the others' weight is removed.

Scale invariance See example in Figure 5. Taking the bottom example and for the group that includes all graphs for which it is a subgraph and those that can be formed by shift invariance and only for them do not apply weakest link but rather the example (it won't trample on the top example, as it can't be reached by shift invariance, and as the edge weights are all minimal/maximal, they change change by ordinal invariance).

Shift invariance See example in Figure 6. Taking the bottom example and for the group that includes all graphs for which it is a subgraph and those that can be formed by scale invariance and only for them do not apply weakest link but rather the example (it won't trample on the top example, as it can't be reached by shift invariance, and as the edge weights are all minimal/maximal, they change change by ordinal invariance).

Monotonicity A strongest link tree implements all axioms but monotonicity.

Path ordinal invariant Minimal spanning tree implements all other axioms.

□

9. DISCUSSION

In this paper we explore the basic issue of routing – how should information flow through a network and what properties might this process have. In the process of considering this issue we developed several properties we believe might be desirable by system planners. For example, *robustness*, or the ability of a routing protocol to keep small changes from disrupting the whole routing process, is a property especially required in fast, changing networks.

Naturally, creating a structure from possible interactions between agents defined by a connections' graph is not limited just to information routing in networks such as the internet. Looking at organizations, where workers are connected according to their ability to work with other workers, and instead of routing messages between them we seek to construct an organizational hierarchy, we face a similar challenge. Again, robustness is a desirable property, as it means that if some workers have a worsening relationship with others, if they're not very senior in the organization, it has little effect on many others. In this case, we may consider the "economic model" ("first hop" axiom) appropriate as well – if workers only interact with their boss, we only care about the edge from each worker to his/her boss, and each worker does not care what happens further up in the hierarchy[2].

Beyond setting up the axioms, we also examined common routing algorithms – minimal spanning tree, shortest path and weakest link, and fully characterized them. Obviously, this is only the beginning of the road for this line of research – further steps will entail developing more axioms and using them to characterize more algorithms, with the aim of giving a set of tools for system designers, allowing them to choose desirable properties which would dictate appropriate routing protocols.

Acknowledgments

The authors thank Michael Schapira for his insightful discussions on this matter. Moshe Tennenholtz carried out this work while at Microsoft Research, Israel. Aviv Zohar is supported in part by the Israel Science Foundation (Grants 616/13 and 1773/13), and by the Israel Smart Grid (ISG) Consortium.

[2] Similarly, in a highly centralized organization, a path cardinal invariance is probably a sensible axiom.

10. REFERENCES

[1] A. Altman and M. Tennenholtz. Ranking systems: The pagerank axioms. In *Proceedings of the 6th ACM conference on Electronic commerce (EC)*, pages 1–8, Vancouver, Canada, June 2005.

[2] A. Altman and M. Tennenholtz. An axiomatic approach to personalized ranking systems. *Journal of the ACM*, 57(4):1–35, 2010.

[3] R. Andersen, C. Borgs, J. Chayes, U. Feige, A. Flaxman, A. Kalai, V. Mirrokni, and M. Tennenholtz. Trust-based recommendation systems: an axiomatic approach. In *Proceedings of the 17th international conference on World Wide Web (WWW)*, pages 199–208, Beijing, China, April 2008.

[4] K. J. Arrow. *Social Choice and Individual Values*. Yale University Press, 1951.

[5] E. Durfee, V. R. Lesser, and D. D. Corkill. Coherent cooperation among communicating problem solvers. *IEEE Transactions on Computers*, 36(2):1275–1291, 1987.

[6] Y. Emek, R. Karidi, M. Tennenholtz, and A. Zohar. Mechanisms for multi-level marketing. In *Proceedings of the 12th ACM conference on Electronic commerce (EC)*, pages 209–218, San Jose, Califronia, June 2011.

[7] M. S. Fox. An organizational view of distributed systems. *IEEE Transactions on Systems, Man and Cybernetics*, 11(1):70–80, 1981.

[8] J. R. Galbraith. *Designing Complex Organizations*. Addison-Wesley, 1973.

[9] L. Gao and J. Rexford. Stable internet routing without global coordination. *IEEE/ACM Transactions on Networking*, 9(6):681–692, December 2001.

[10] T. G. Griffin and J. L. Sobrinho. Metarouting. In *Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM)*, pages 1–12, Philadelphia, Pennsylvania, August 2005.

[11] S. Herzog, S. Shenker, and D. Estrin. Sharing the ”cost” of multicast trees: An axiomatic analysis. *IEEE/ACM Transactions on Networking*, 5(6):847–860, December 1997.

[12] M. Karsten, S. Keshav, S. Prasad, and M. Beg. An axiomatic basis for communication. In *Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM)*, pages 217–228, Kyoto, Japan, August 2007.

[13] D. Kotz, C. Newport, R. S. Gray, J. Liu, Y. Yuan, and C. Elliott. Experimental evaluation of wireless simulation assumptions. In *Proceedings of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems (MSWiM)*, pages 78–82, Venice, Italy, October 2004.

[14] H. Levin, M. Schapira, and A. Zohar. Interdomain routing and games. In *Proceedings of the 40th annual ACM symposium on Theory of computing (STOC)*, pages 57–66, Victoria, Canada, May 2008.

[15] T. W. Malone. Informational efficiency in networks and hierarchies. Sloan School of Management Working Paper 1849, MIT, 1986.

[16] J. G. March and H. A. Simon. *Organizations*. John Wiley and Sons, 1958.

[17] Y. L. Sun, W. Yu, Z. Han, and K. J. Liu. Information theoretic framework of trust modeling and evaluation for ad hoc networks. *IEEE Journal on Selected Areas in Communications*, 24(2):305–317, September 2006.

[18] T. S. Verma and J. Pearl. Causal networks: Semantics and expressiveness. In *Proceedings of the 4th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 69–78, 1988.

Preference at First Sight

Chanjuan Liu
School of Electronics Engineering and Computer Science, Peking University
Institute for Logic, Language and Computation, University of Amsterdam
chanjuan.pkucs@gmail.com

ABSTRACT

We consider decision-making and game scenarios in which an agent is limited by his/her computational ability to foresee all the available moves towards the future – that is, we study scenarios with *short sight*. We focus on how short sight affects the logical properties of decision making in multi-agent settings. We start with *single-agent sequential decision making* (SSDM) processes, modeling them by a new structure of '*preference-sight trees*'. Using this model, we first explore the relation between a new natural solution concept of *Sight-Compatible Backward Induction* (SCBI) and the histories produced by classical Backward Induction (BI). In particular, we find necessary and sufficient conditions for the two analyses to be equivalent. Next, we study how computational complexity changes when short-sight is involved, and also, whether computationally costly larger sight always contributes to better outcomes. Then we develop a simple logical special-purpose language to formally express some key properties of our preference-sight models. Lastly, we show how short-sight SSDM scenarios call for substantial enrichments of existing fixed-point logics that have been developed for the classical BI solution concept. We also discuss changes in earlier modal logics expressing 'surface reasoning' about best actions in the presence of short sight. Our analysis may point the way to logical and computational analysis of more realistic game models.

1. INTRODUCTION

There is a growing interest in the logical foundations, computational implementations, and practical applications of *single-agent sequential decision-making* (SSDM) problems [13; 18; 9; 16; 14] in such diverse areas as Artificial Intelligence, Control, Logic, Economics, Mathematics, Politics, Psychology, Philosophy, and Medicine. Making decisions is central to agents' routine and usually, they need to make multiple decisions over time. Indeed, a current situation is a result of past sequentially linked decisions, each impacted by the preceding choices.

It is quite natural in sequential decision-making scenarios, particularly, in large systems, that agents may have some uncertainties and limitations on their precise view of the environment. The current literature [18] has studied uncertainty which an agent faces in recognizing possible outcomes after taking an action and the probabilities associated with these outcomes, as well as the partial observability of what the actual state is like. In addition to these, a realistic aspect that affects a SSDM process is the short-sightedness of the agent, which blocks a full view of all the available actions. Short sight plays a critical role in such a situation, since, while making a choice, the ability to foresee a variety of alternatives and predict future decision sequences for each of them, may make a significant difference. Nonetheless, such restrictions have not been discussed systematically yet in decision theory or game theory.

In [6], a game-theoretic framework called *games with short sight* was proposed. This framework explicitly models players's limited foresight in extensive games and calls for a new solution termed as *Sight-Compatible Backward Induction* (SCBI). However, many essential issues related to sight remain unclear, such as: What is the exact role of sight? Will the outcome be better when sight is larger? What is the relation between SCBI and classical *backward induction*(BI)? There are also unexplored issues pertaining to logical aspects. Which minimal logic is needed for formally characterizing a short-sight framework? Are existing logics for BI still applicable, or can they be extended to fit short-sight scenarios? How different are the logical properties of the game frames for SCBI and for BI? Without such a logical analysis, the framework of [6] does not suffice for disclosing the general features of short sight and the changes it brings about in thinking about decisions and games. Additionally, in multi-player games, short sight has to interact with many other factors, such as agents' mutual knowledge and interactive decisions and moves.

Having said this, we still start by focusing on short sight in single-agent sequential decision-making process. For this, we propose a model of '*preference-sight trees*' (P-S trees). As the term says, a P-S tree combines the agent's *preference* and its sight, as both are essential to decision problems [21]. We will study how the two are correlated, and cooperate to act on decision-making processes and their final outcomes.

As a preliminary illustration, consider the connection between larger sight and better outcome. A first impression might be that an agent will always perform better with larger sight. Surprisingly, this is not always true. Sometimes, one can see much further into the future but receive a small payoff, while having one's vision restricted to a limited set of future alternatives yields a better payoff.

EXAMPLE 1.1. *Alice has to make sequential decisions at two stages (shown in Figure 1). For each stage, she can choose either L or R. Assume that the preference order (from most preferable to least preferable) among the four outcomes is RR, LL, RL, LR. Now consider two cases:*

Case 1. At the start, Alice sees two paths, LR and RL. She chooses R since it initiates RL which is preferable to LR. At the second-stage, Alice then foresees RR and RL.

She happily makes the best decision RR.

Case 2. Alice sees more, e.g., LL, LR, and RL, immediately at the first stage. Therefore she thinks that L is a better initial choice than R. Consequently, at the second-stage, she can only choose from LL and LR.

Conclusion: Even though Alice could see more in Case 2, she ultimately obtains a less preferable outcome.

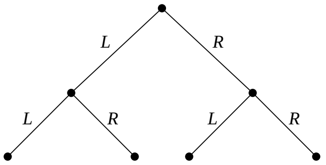


Figure 1: Two-stage decision-making

This example demonstrates some of the crucial features that govern SSDM situations:

1) What an agent can foresee plays a crucial role in the decision-making process, since her sight determines the set of available choices.

2) Sight also updates her preferences over the options, and thereby the outcomes obtained in rational play.

3) Although in Case 2, *Alice* does not get the best result, we can say that, given her sight, she plays optimally in a local sense. In other words, this is a rational plan for her, even though it is not equivalent to the rational outcome of classical decision theory or game theory [19].

In this paper, we address all three challenges, but first we clarify our approach. To focus on sight, we ignore other factors such as the probability of moves by Nature. Also, we model the outcome of a decision as completely determined, or in other words, possible outcomes for each alternative and the probability corresponding to each outcome are encapsulated as a black box.

2. MODELING SINGLE-AGENT SEQUENTIAL DECISION-MAKING

We begin by defining a structure called *preference-sight tree* for modelling single-agent sequential decision-making (SSDM) processes. Using this model, we then clarify the role that sight plays by discussing a series of changes it produces in agent's preferences, decision-making procedures and their outcomes, as well as computational complexity.

2.1 Models

There are two kinds of models for decision-making scenarios corresponding to two perspectives. One is an *explicit* model from the perspective of Nature, or an outsider/designer; the other is the *implicit* model from the perspective of the agent involved, or an insider/decider. The former is complete and perfect in the sense that the outsider holds a full view of all the options together with the *objective* quality of these options, and thus can explicitly specify the reward of each situation for the decision-maker. In contrast with this, the latter's views are possibly limited to a near future, especially in large-scale surroundings. Moreover, owing to limited foresight, the agent may also reason mistakenly about the quality of different choices, leading to what we call *subjective* preference.

Both the above perspectives are essential: the former offers a whole picture of the environment, the latter shows the actual play of the decider. In this section, we first introduce an explicit model of *preference trees.* After this, by endowing such trees with the agent's view of the process and his/her subjective preference in this view, we formulate an integrated model of *preference-sight trees* which allows us to model both perspectives together.

2.1.1 Preference trees (P trees)

A preference tree is a decision tree with only two elements: histories and preferences. Each history corresponds to a situation resulting from previous decision actions, and a preference represents the objective quality of each of these situations. To ensure the existence of backward induction solutions, we confine ourselves to finite histories.

DEFINITION 2.1. (*Preference tree*) *A* **preference tree** *is a tuple $T = (H, \succeq)$ where H is a non-empty set of finite sequences of actions, called* histories; *$\succeq$ is a total order over H. The empty sequence ε is a member of H; If $(a^k)_{k=1,\dots,K} \in H$ and $L < K$ then $(a^k)_{k=1,\dots,L} \in H$.*

Let A denote the set of all actions. Any history h can be written as a sequence of actions: $(a^k)_{k=1,\dots,n}$, where each $a^k \in A$. If there is no a^{n+1} s.t. $(a^k)_{k=1,\dots,n+1} \in H$, then history $(a^k)_{k=1,\dots,n}$ is a terminal one. The set of terminal histories is denoted Z. The set of actions that are available at h is denoted $A(h) \subseteq A$. For any histories h, h', if h is a prefix of h' we write $h \lhd h'$. The strict part of $\succeq$ is $\succ$, with $h_1 \succ h_2$ if $h_1 \succeq h_2$ and not $h_2 \succeq h_1$ for any two histories h_1 and h_2. Accordingly, $h_1 \sim h_2$ iff $h_1 \succeq h_2$ and $h_2 \succeq h_1$.

Several remarks need to be made on the role of preference relations in the above definition:

(1) Instead of defining preference merely over terminal histories, we have defined it over all histories, an idea going back to [11]. Here preference over intermediate histories is necessary for our aim of modelling an agent's decision-making under limited foresight, which usually consists of intermediate histories.

(2) For convenience, we do not strictly differentiate the two main views of preference: qualitative and quantitative. Although we use qualitative order generally, we sometimes switch to numerical payoff when it is advantageous.[1]

2.1.2 Preference-sight Trees (P-S trees)

P tree is an explicit model for decision-making scenarios which is independent of an agent. However, for an agent, the tree may appear differently in his/her limited view. [6] proposes the idea of short sight, where the authors use a sight function to denote the set of states that players can actually see at every position in an extensive game. Let us start by adapting their technique to preference trees.

DEFINITION 2.2. *Let $T = (H, \succeq)$ be a preference tree. A* **sight function** *for T is a function $s : H \to 2^H \backslash \{\emptyset\}$ satisfying $s(h) \subseteq H|_h$ and $|s(h)| < \omega$, where $H|_h$ represents the set of histories extending h. As a special case, $h \in H|_h$.*

In words, the function s assigns to each history h a finite subset of all available histories extending h.

The first effect that sight produces is that given a P tree, for any history h, it always gives us a restricted tree.

[1] There is a debate on whether preference and utilities are the same [9; 2]. Here we adopt the operational understanding of utility and do not distinguish it from preference.

Definition 2.3. *Let $T = (H, \succeq)$ be a P tree. Given any history h of T, a* **visible tree** *T_h of T at h is a tuple $(H_h, \succeq_h)$, where $H_h = s(h)$, i.e., H_h captures the decider's view of the decision tree; $\succeq_h$ represents the subjective preference over H_h.*

A visible tree is actually an implicit model in our earlier terms. H_h also contains a set of terminal histories Z_h, which are those without successors in $s(h)$. Note that typically, the Z_h are non-terminal for T.

Further, the preference order $\succeq_h$ is different from the objective preference $\succeq$. In fact, the formation of $\succeq_h$ is an update via a bottom-to-top process in terms of an agent's sight. This updating process involves leaving the payoffs of Z_h as the same as their objective payoffs, then updating the payoffs of other histories in H_h backwards, starting from the leaf nodes and proceeding towards the root of the tree.

The reason why we employ such an updating process is that, while the objective payoffs reflect the goodness of these situations, they are not the actual reward that an agent can get if he/she chooses this option. At each decision point, the subjective payoff of one available option is inherited from the best reachable terminal histories of the current visible tree. Therefore, the preference relation $\succeq_h$ in T_h is not always consistent with the preference relation $\succeq$ in T.

This updating process is described by Algorithm 1, which essentially involves a backward computation and update of the preference over intermediate nodes within the sight:

*For convenience, here we use payoffs P to represent rewards.

Algorithm 1: Preference updating in visible trees

```
PU($T, h, s$)
Input: A P tree $T = (H, \succeq)$ (or $T = (H, P)$), current
       history $h$, and a sight function $s$
Output: A visible tree $T_h = (H_h, \succeq_h)$ or
        ($T_h = (H_h, P_h)$)
begin
    $H \cap s(h) \to H_h$;
    for any $z \in Z_h$ /* Keep the payoffs of terminal
    histories unchanged */ do
        $P(z) \to P_h(z)$;  $1 \to flag[z]$;
    while $flag[h] == 0$ do
        for any $h' \in H_h$ do
            if (for all $(h'a) \in H_h$, $flag[(h'a)] == 1$)
            /* If all of its children have been
            visited, reset its payoff as the
            highest one among them */
            then
                $max\{P_h(h'a)\} \to P_h(h')$;  $1 \to flag[h']$;
    Return $T_h$;
```

Fact 2.1. *Let $T = (H, \succeq)$ be a P tree. Each visible tree $T_h = (H_h, \succeq_h)$ is a P tree.*

Correspondingly, we denote the prefix relation in T_h by $\lhd_h$, and the actions that are available at h by $A_h(h)$.

Finally we proceed to define our model of preference-sight trees. A preference-sight tree allows us not only to represent the outsider's view, i.e., $(H, \succeq)$, but also to derive a series of implicit models, i.e., $(H_h, \succeq_h)$, one for each h.

Definition 2.4. (*Preference-sight tree*) *A* **preference-sight tree** *(P-S tree) is a tuple (T, s), where $T = (H, \succeq)$ is a preference tree and s a sight function for T.*

In P-S trees, an agent's sight should satisfy the following properties: First, if an agent can see a given future history, then he/she can also see any intermediate history up to that point. Second, if the agent can see a history two steps forward, then after moving one step ahead, he/she can still see it. These features are formally stated as follows.

Fact 2.2. (*Properties of sight function*) *Let (T, s) be a P-S tree. For all $h, h', h'' \in H$, with $h \lhd h' \lhd h''$, s satisfies :*

DC *(Downward-Closed): if $h'' \in s(h)$, then $h' \in s(h)$.*

NF *(Non-Forgetting): if $h'' \in s(h)$, then $h'' \in s(h')$.*

2.2 Solution concepts

Solution concepts are at the center of all choice problems. In what follows, we define two solution concepts for P-S trees, adapted from [20; 6]. After this, we investigate the conditions for their equivalence, followed by providing procedures for calculating the number of them.

2.2.1 BI history and SCBI history

Backward Induction (BI) is a well-known process running like this. First, one determines the optimal strategy of the player who makes the last move of the game. Using this information, one can then determine the optimal action of the next-to-last moving player. The process continues backwards in this way until all players' actions have been determined in the whole game. Its adaptation to single-agent decision-making process becomes a maximality problem for the agent involved.

In a P-S tree, we say that one history h is $max_{\succeq}$ in a set of histories $\Gamma \subseteq H$, if $h \in \Gamma$ and for any other history h' in Γ, it holds that $h \succeq h'$, and we write this as $h \in max_{\succeq}\Gamma$. The strict part for $max_{\succeq}$ is $max_{\succ}$.

Definition 2.5. (BI *history*) *Let (T, s) be a P-S tree. A history $h^* \in Z$ is a* **BI history** *of T, iff $h^* \in \max_{\succeq} Z$. Also, we use* **BI** *to denote the set of BI histories in T.*

A BI history of a P-S tree is a terminal history that is most preferable or equivalently, that has a maximal payoff.

Backward induction precludes short-sight, while in practice it is impossible for an agent to foresee all final outcomes all the time. In [6], a new solution concept was proposed to capture optimal play of short-sighted players: *sight-compatible subgame perfect equilibrium.* The main idea is that at each decision point, the current player chooses a locally optimal move by a local BI analysis within the visible part. Here, we adapt this notion to P-S trees, yielding the *sight-compatible backward induction history.*

Definition 2.6. (SCBI *history*) *Let (T, s) be a P-S tree. A history $h^* \in Z$ is a* **Sight-Compatible Backward Induction history** (SCBI history) *of T, iff for each history h with $h \lhd h^*$, and the action a following h, i.e., $(ha) \lhd h^*$, we have that $\exists z \in \max_{\succeq} Z_h$ such that $(ha) \lhd z$. Also, we use* **SCBI** *to denote the set of SCBI histories in T.*

The difference between SCBI and BI histories is obvious. A BI history is one with highest payoff among the set of terminal histories in the P-S tree, while for a SCBI history every restriction of it should be a local BI history for the visible tree. Thus, BI histories are the BI outcomes for the objective model $(H, \succeq)$, while SCBI histories are a combination of best responses to all subjective models $(H_h, \succeq_h)$. Typically it is the case that $\mathbf{SCBI} \neq \mathbf{BI}$.

Example 2.1. *Consider the P-S tree (T, s) in Figure 2, where $s(\varepsilon) = \{L\}$, and $s(L) = \{LR\}$. It is easy to check that $\mathbf{BI} \neq \mathbf{SCBI}$, since $\mathbf{BI} = \{LL\}$, while $\mathbf{SCBI} = \{LR\}$.*

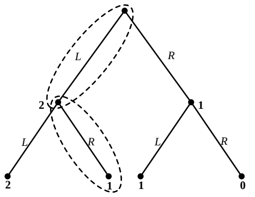


Figure 2: $\mathbf{BI} \neq \mathbf{SCBI}$

However, sometimes the two notions can be equivalent.

Example 2.2. *Consider a P-S tree, with T and s shown by Figure 3 (a), and Figure 3 (b) respectively. In (b) the three dotted circles represent $s(\varepsilon)$, $s(L)$ and $s(R)$. For histories L and R, their objective payoffs in (a) are* 1 *and* 2*, respectively. However, in T_ε, the subjective payoff of L is updated to* 3 *and R to* 2*. Obviously, $\mathbf{BI} = \mathbf{SCBI} = \{LL\}$.*

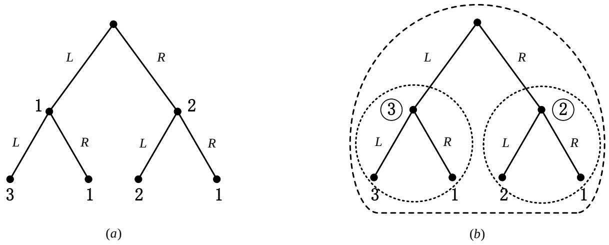


Figure 3: ($\mathbf{BI} = \mathbf{SCBI}$)

2.2.2 Equivalence condition

Then an interesting question on BI and SCBI histories arises: are there conditions under which the two will be equivalent? To get a feeling for this, a first attempt at an answer looks for a condition related to consistency between subjective and objective preferences.

Two histories are said to be 'preference-sight consistent' if the subjective preference in each sight-restricted tree is consistent with the objective preference over them:

Definition 2.7. (*Preference-sight consistency*) *Let (T, s) be a P-S tree, and T_h be the visible tree at an arbitrary history h. Then for any two histories h_1, h_2 of T_h, we say (h_1, h_2) satisfies* **preference-sight consistency** *at h iff*

$$h_1 \succeq h_2 \text{ iff } h_1 \succeq_h h_2$$

If for any history $h \in T$, the pair of arbitrary two histories (h_1, h_2) in T_h is *preference-sight consistent* (at h), then we say (T, s) is *preference-sight consistent*.

Is preference-sight consistency an appropriate condition for $\mathbf{BI} = \mathbf{SCBI}$? We have the following observation:

Fact 2.3. *Preference-sight consistency does not guarantee that* $\mathbf{BI} = \mathbf{SCBI}$.

Proof. Consider Figure 2. Suppose that $s(R)$ contains only one successor. Then it is easy to see that (T, s) is preference-sight consistent. However, $\mathbf{BI} \neq \mathbf{SCBI}$. □

Next, does the other direction hold?

Fact 2.4. *Preference-sight consistency does not follow from* $\mathbf{BI} = \mathbf{SCBI}$.

Proof. The situation in Figure 3 is a counterexample, in which $\mathbf{BI} = \mathbf{SCBI} = \{LL\}$, but (T, s) is not preference-sight consistent, since $R \succ L$ and $L \succ_\varepsilon R$. □

What is the exact condition for $\mathbf{BI} = \mathbf{SCBI}$? From the failure of preference-sight consistency, we can draw a lesson. In Figure 2, the main reason for (T, s) being inconsistent is that at history L, the branch LL, which in fact forms a BI history, is non-observable to the agent. This tells us that the one with maximal payoff should always be visible. Consider then the example in Figure 3. Here all the options are within agent's sight, but we notice that although the path LL following L finally turns out to be better than that following R, which makes subjectively $L \succ_\varepsilon R$, the objective payoff of L itself is lower than R. Thus, it fails to imply the consistency between preference and sight.

Based on the above analysis, we now isolate necessary and sufficient conditions for $\mathbf{BI} = \mathbf{SCBI}$. First, we define an auxiliary property of *sight-reachability*, which intuitively reflects whether each restriction of a history is visible.

Definition 2.8. (*Sight-reachability*) *A BI history h^* is* **sight-reachable** *if, for all $(ha) \lhd h^*$, we have $(ha) \in H_h$, where h, h' are histories, and a is an action following h.*

Theorem 2.5. (*Equivalence Theorem*) *For any P-S tree (T, s),* $\mathbf{SCBI} = \mathbf{BI}$ *iff the following conditions are satisfied:*

I). *Any history $h^* \in \mathbf{BI}$ is sight-reachable.*

II). *Any history $h^* \in \mathbf{BI}$ is locally optimal: For any history $(hh') \lhd h^*$, if $(hh') \in Z_h$, then $(hh') \in \max_{\succ} Z_h$ and for any other $(hh'') \in Z_h$, $(hh') \sim (hh'')$ iff $\exists z \in \mathbf{BI}$ such that $(hh'') \lhd z$.*

Proof. ($\Rightarrow$) I). We show that every $h^* \in \mathbf{BI}$ is sight reachable. That is, for all $(hh') \lhd h^*$, it holds that $(ha) \in H_h$. By $\mathbf{SCBI} = \mathbf{BI}$, we know that any history h^* in $\mathbf{BI}$, is also in $\mathbf{SCBI}$. By Definition 2.6, for each of its prefix h, h^*_h is $max_{\succeq}$ in Z_h. So h^*_h is in Z_h. In addition, by non-emptiness of Z_h, h^*_h is not an empty sequence. Thus, for all $(ha) \lhd h^*$, it holds that $(ha) \in H_h$. So $h^* \in \mathbf{BI}$ is sight-reachable.

To show condition II), take any h^* in $\mathbf{BI}$, we have that it is in $\mathbf{SCBI}$. Thus, for all $(hh') \lhd h^*$, if $(hh') \in Z_h$, then (hh') is $max_{\succeq}$ in Z_h. Moreover, for any $(hu) \in Z_h$ such that $(hh') \sim (hu)$, we have (hu) is a prefix of a BI history, i.e., $(hu) \in \mathbf{BI}_h$. For suppose not, then (hu) is not a prefix of SCBI history. Then it must be $(hh') \nsim (hu)$. Contradict.

($\Leftarrow$) Suppose conditions I) and II) are satisfied. It suffices to show (a)"every BI history is SCBI history of T", and (b) " every SCBI history is BI history of T".

For (a), take any BI history h^*. By I), all BI histories are sight reachable. Further by II), for all $(hh') \lhd h^*$, if $(hh') \in Z_h$, then (hh') is $max_{\succeq}$ in Z_h. This is to say that for each of its prefix h, h^*_h is $max_{\succeq}$ in Z_h. By definition 2.6, h^* is a SCBI history.

For (b), take any SCBI history h^*. We can show it is a BI history, i.e., h^* is $max_{\succeq}$ in Z. For suppose not, then there exists a BI history h' such that $h' \succ h^*$.

Notice that there must be some history u which is the common prefix of h^* and h'. Since h' is a BI history, by condition I) and II), we know that $h'_u \succ h^*_u$. Then h^*_u is not a prefix of a SCBI history. Thus, h^* is not a SCBI history. Contradiction. □

2.2.3 More sight, better outcome?

We have seen earlier on that, SCBI may loss global optimality. The BI history definitely has a maximal payoff, while it might not be the case for SCBI, since each action is chosen with a limited sight. So **BI** $\succeq$ **SCBI** holds without exception, in the sense that any BI history is no worse than any SCBI history. One might conjecture that more sight always contributes to better outcomes. Yet, the fact below falsifies this.

FACT 2.6. *Let T be a P tree. Also, let s_1 and s_2 be two sight functions for T satisfying $s_1(h) \subseteq s_2(h)$ for any history h in T. Take any two* SCBI *histories z_1 and z_2 of (T, s_1) and (T, s_2) respectively. Then the following three cases are all possible: a) $z_1 \succ z_2$; b) $z_2 \succ z_1$; c) $z_1 \sim z_2$.*

PROOF. Case (a) has been shown in Example 1.1. Case (b): Obviously, Figure 2 offers an instance for this. Case (c): The scenario depicted in Figure 3 is an example. □

In conclusion, full sight guarantees a maximal payoff. However, with short sight, increase of sight does not always improve the outcome. The added sight may bring misleading information, e.g., a branch which is temporarily nicer but actually unpromising, and finally gives rise to an even worse outcome. Still, this does not mean that SCBI is deficient: rather, these observations seem realistic for real agents. These issues will be discussed further in Section 4.

3. A LOGICAL ANALYSIS

After modelling decision-making with short sight by preference tree models, it is instructive to see what a logical language looks like for reasoning about these models, especially the role of sight in a SSDM process. So far, no such logic has been proposed, though logics of game-theoretic structures have been extensively studied – see [27; 10] – while there are a few preliminary logic analyses of sight on its own, [4; 17]. In this section, we design a minimal and natural logical system that supports reasoning about sight in the context of single-agent decision-making processes, characterizing basic properties of preference-sight trees, and formally capturing the results in the previous section.

3.1 Syntax and Semantics

To reason about the key ingredients (i.e., histories, preferences, and sights) of a P-S tree, we take $P^{(T,s)}$ as a set of propositional letters, which at least contains the following [2]:

- $\overline{h}$ for each history h.
- $\overline{h_1 \geq h_2}$ encoding the preference relation of the agent over all histories, and the strict part of which is $\overline{h_1 > h_2}$.
- $\overline{s(h)}$ encoding the sight at each history h in T.

Based on $P^{(T,s)}$, we give a language $\mathcal{L}$ for reasoning about P-S trees. In $\mathcal{L}$, we have a key dynamic operator $[!\varphi]$ for restricting to the worlds satisfying φ, and a universal modality with $A\varphi$ saying that φ is true in every world.

DEFINITION 3.1. (*Preference-sight language*) *Take any set of atomic letters $P^{(T,s)}$. The* **preference-sight language** *$\mathcal{L}$ is given by the following BNF, where $p \in P^{(T,s)}$:*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid [!\varphi]\psi \mid A\varphi.$$

We write $\langle !\varphi \rangle \varphi$ to abbreviate $\neg[!\varphi]\neg\varphi$.

DEFINITION 3.2. (*Preference-sight models*) *For a P-S tree (T, s), a* **preference-sight model** *$M^{(T,s)}$ is a tuple $(H, \lhd, \mathcal{V})$ where the following holds:*

- *H is the set of possible worlds, one for each history,*
- *$\lhd$ is the reachability (prefix) relation among worlds,*
- *$\mathcal{V} : P_T \to \rho(H)$ is an evaluation function satisfying:*
  (1) $\forall h \in H$, $\mathcal{V}(\overline{h}) = \{h' | h' \lhd h\}$.
  (2) $\mathcal{V}(\overline{h_1 \geq h_2}) = \begin{cases} H, & \text{IF } h_1 \succeq h_2, \\ \emptyset, & \textit{Otherwise.} \end{cases}$
  (3) $\forall h \in H$, $\mathcal{V}(\overline{s(h)}) = \bigcup_{h' \in s(h)} \mathcal{V}(\overline{h'})$.

Intuitively, $\overline{h}$ is true at all the worlds leading to h. $\overline{h_1 \geq h_2}$ is true everywhere if $h_1 \succeq h_2$, and nowhere otherwise. Finally, $\mathcal{V}(\overline{s(h)})$ is a union of the worlds that make the given atom true for at least one element of $s(h)$.

There seems to be nothing striking in this syntax. However, given the special role of atoms, the natural model update differs from the usual one in dynamic-epistemic logic.

DEFINITION 3.3. *(Model update) Given a preference-sight model $M^{(T,s)} = (H, \lhd, \mathcal{V})$ and a set $X \subset H$, the* ***updated model*** *$M^{(T,s)}_{!X}$ produced by the restriction of X is defined as a tuple $(X, \lhd \cap X^2, \mathcal{V}_{!X})$, where* [3]

$$\mathcal{V}_{!X}(p) = \begin{cases} \mathcal{V}_{!X}(\overline{h_1 \geq h_2}), & \textit{IF } p \textit{ is of the form } \overline{h_1 \geq h_2} \\ \mathcal{V}(p) \cap X, & \textit{Otherwise} \end{cases}$$

$$\mathcal{V}_{!X}(\overline{h_1 \geq h_2}) = \begin{cases} X, & \textit{IF } \mathcal{V}(\overline{z_1 \geq z_2}) = H, \textit{ where} \\ & \quad z_1 \in max_{\succeq}\{z \in Z_X | h_1 \lhd z\}, \\ & \quad z_2 \in max_{\succeq}\{z \in Z_X | h_2 \lhd z\} \\ \emptyset, & \textit{Otherwise} \end{cases}$$

$M^{(T,s)}_{!X}$ is the update of the model $M^{(T,s)}$ restricting the set of states to X, and the valuation function accordingly. But crucially, the valuation for preference atoms in the new model reflects the updating process in the visible tree of Algorithm 1. In the following, we omit superscripts (T, s).

The semantics for this language is basically standard, [3], so we only mention the truth condition of $[!\varphi]\psi$:

Let M be a preference-sight model. For any state h in M,

$$M, h \models [!\varphi]\psi \textit{ iff } M, h \models \varphi \Rightarrow M_{!\llbracket\varphi\rrbracket}, h \models \psi,$$
$$\textit{where } \llbracket\varphi\rrbracket = \{h' \in H | M, h' \models \varphi\}.$$

Validity of formulas is defined as usual, cf. [3].

[2] The idea of defining $\overline{h}$ is motivated by [1], where the authors define an atomic sentence $\overline{o}$ for each leaf in a game tree.

[3] In this definition, Z_X denotes the terminal histories in X, i.e., the set of histories that have no successors in X.

3.2 Main characterization results

Despite its simplicity, $\mathcal{L}$ can express our results in previous sections concerning properties and solutions of P-S trees. We introduce some helpful syntactic abbreviations, and then state our main characterization results.

- $\overline{Z_h} = \bigvee\{\ \overline{z} \mid z \in Z_h\}$.
- $\overline{max_{\geq} X} {=} \bigvee\{\ \overline{h} \mid h \in X,\ \textit{and}\ h \succeq h'\ \textit{for}\ \forall h' \in X\}$.
- $\overline{\mathrm{BI}} = \bigvee\{\ \overline{z} \mid z \in \mathbf{BI}\}$ ($\overline{\mathrm{BI}}$ holds at T's BI histories).
- $\overline{\mathrm{SCBI}} = \bigvee\{\ \overline{z} \mid z \in \mathbf{SCBI}\}$, that is, the formula $\overline{\mathrm{SCBI}}$ holds at the SCBI histories of T.

Proposition 3.1. *Let (T, s) be a P-S tree and M be a $\mathcal{L}$-model for it. Then (T, s) is* preference-sight consistent *iff the following formula is valid in M:*

$$\bigwedge_{h} \bigwedge_{h_1 \in H_h} \bigwedge_{h_2 \in H_h} ((\overline{h_1 \geq h_2} \rightarrow [!\overline{s(h)}]\overline{h_1 \geq h_2}) \wedge$$
$$(\langle !\overline{s(h)} \rangle \overline{h_1 \geq h_2} \rightarrow \overline{h_1 \geq h_2})).$$

Lemma 3.2. *For any P-S tree (T, s) and model M for it, a BI history h^* is* sight-reachable *if and only if the following formula holds in M:*

$$(SR): \quad \bigwedge_{h} \bigwedge_{a \in A(h)} (A(\overline{(ha)} \rightarrow \overline{h^*}) \rightarrow (A(\overline{(ha)} \rightarrow \overline{s(h)}))).$$

Proof. ($\Rightarrow$) Suppose that BI history h^* is sight-reachable. By Definition 2.8, we have that, for all $(ha) \lhd h^*$, it holds that $(ha) \in s(h)$, where h, h' are histories, and a is an action following h. More formally, $(ha) \lhd h^*$ can be defined by the formula $A(\overline{(ha)} \rightarrow \overline{h^*})$ in the sense that, in T, for all h and $a \in A(h)$, $(ha) \lhd h^*$ iff $M \models A(\overline{(ha)} \rightarrow \overline{h^*})$. And similarly $(ha) \in s(h)$ is defined by $A(\overline{(ha)} \rightarrow \overline{s(h)})$. Thus if a BI history h^* is sight-reachable, then $M \models \bigwedge_h \bigwedge_{a \in A(h)} (A(\overline{(ha)} \rightarrow \overline{h^*}) \rightarrow (A(\overline{(ha)} \rightarrow \overline{s(h)})))$. The other direction can be proved in a similar way. □

Lemma 3.3. *Let (T, s) be a P-S tree and M be a $\mathcal{L}$-model for it. A* BI *history h^* is* locally optimal *iff the following formula is valid in M:*

$$(LO): \quad (\bigwedge_{h} \bigwedge_{(hh') \in Z_h} (A(\overline{(hh')} \rightarrow \overline{h^*}) \rightarrow$$
$$(A(\overline{(hh')} \rightarrow \overline{\max_{\succeq} Z_h}) \wedge$$
$$\bigwedge_{(hh'') \in Z_h} (\overline{(hh') \sim (hh'')} \leftrightarrow \bigvee_{z \in \boldsymbol{BI}} (A(\overline{(hh'')} \rightarrow \overline{z}))))).$$

Proof. ($\Leftarrow$) Suppose BI history h^* is locally optimal. Then for $(hh') \lhd h^*$, if $(hh') \in Z_h$, we have (hh') is $max_{\succeq}$ in Z_h. And for any (hh''), $(hh'') \sim (hh')$ iff $\exists z \in \mathbf{BI}$ s.t. $(hh'') \lhd z$. Similar with the above proposition, $A(\overline{(hh')} \rightarrow \overline{h^*})$ captures that $(hh') \lhd h^*$. And $A(\overline{(hh'')} \rightarrow \overline{z})$ shows that $(hh'') \lhd z$. Finally, $(\overline{(hh')} \rightarrow \overline{max_{\succeq} Z_h})$ demonstrates that (hh') is $max_{\succeq}$ in Z_h. Direction ($\Rightarrow$) uses a similar check. □

Proposition 3.4. *($\mathcal{L}$-characterization of equivalence) Let (T, s) be a preference-sight tree and M a model for it. Then the following formula is valid in M:*

$$\models (A(\overline{\mathrm{BI}} \leftrightarrow \overline{\mathrm{SCBI}})) \leftrightarrow$$
$$\bigwedge_{h^* \in Z} ((A(\overline{h^*} \rightarrow \overline{\mathrm{BI}})) \rightarrow (SR \wedge LO)).$$

Proof. Direction ($\Rightarrow$). We need to prove the following:

1) $(A(\overline{\mathrm{BI}} \leftrightarrow \overline{\mathrm{SCBI}})) \rightarrow \bigwedge_{h^* \in Z} (A(\overline{h^*} \rightarrow \overline{\mathrm{BI}}) \rightarrow SR)$.

2) $(A(\overline{\mathrm{BI}} \leftrightarrow \overline{\mathrm{SCBI}})) \rightarrow \bigwedge_{h^* \in Z} (A(\overline{h^*} \rightarrow \overline{\mathrm{BI}}) \rightarrow LO)$.

For 1). It is equivalent to prove that, for any $h^* \in Z$, $(\overline{\mathrm{BI}} \leftrightarrow \overline{\mathrm{SCBI}}) \wedge (A(\overline{h^*} \rightarrow \overline{\mathrm{BI}})) \rightarrow SR$. Suppose $\neg(SR)$. Then $\exists (ha) \lhd h^*$, and $(ha) \notin T_h$, and so, at h, the branch leading to h^* is not visible in T_h. Thus, the BI history in T_h could not be a branch leading to h^*. By the definition **SCBI**, it follows that $h^* \notin \mathbf{SCBI}$. However, by $\overline{h^*} \rightarrow \overline{\mathrm{BI}}$ we know that h^* is a BI history. This contradicts $\overline{\mathrm{BI}} \leftrightarrow \overline{\mathrm{SCBI}}$.

2) can be proved in a similar style.

Direction ($\Leftarrow$). Suppose that $\neg(A(\overline{\mathrm{BI}} \leftrightarrow \overline{\mathrm{SCBI}}))$. Then

(a): $\exists z^* \in \mathbf{BI}$ and $z^* \notin \mathbf{SCBI}$, or

(b) : $\exists z^* \in \mathbf{SCBI}$ and $z^* \notin \mathbf{BI}$.

If (a), then, by the antecedent, we have that: $\forall (ha) \lhd z^*, (ha) \in H_h$. Also, $\forall (hh') \in Z_h$ and $(hh') \lhd h^*$, it holds that $(hh') \in max_{\succeq} Z_h$. Then it directly follows that z^* is a SCBI history. Contradiction.

If (b), then take any $z \in \mathbf{BI}$, which shares a prefix u with z^*, i.e., $u \lhd z$ and $u \lhd z^*$. By the antecedent, we have $z_u \in max_{\succeq} Z_h$. Since $z^* \notin \mathbf{BI}$, it follows that $z_u > z^*_u$. Then $z^* \notin \mathbf{SCBI}$. Once more, we have a contradiction. □

3.3 Valid principles

The operator $[!\varphi]$ makes $\mathcal{L}$ a PAL-like language. However, the special model-update makes it different from standard PAL [28]. This suggests a close look at what is and what is not valid in preference-sight models.

First, some axioms in standard PAL do not hold in preference-sight models. For example, the $!ATOM$ axiom, $[!\varphi]p \leftrightarrow (\varphi \rightarrow p)$, is not valid when it is of the form below.

Proposition 3.5. *The following is not valid in preference-sight models, where h, h_1, h_2 represent arbitrary histories.*

$$!\text{Sight-Preference}: \quad [!\overline{s(h)}]\overline{h_1 \geq h_2} \leftrightarrow (\overline{s(h)} \rightarrow \overline{h_1 \geq h_2}).$$

This proposition says that subjective preference in visible trees is not necessarily consistent with objective preference.

Now let us see some interesting valid principles and their intuitive interpretations.

Lemma 3.6. *The formulas shown in Table 1 are valid, where h, h_1, h_2, and h_3 are arbitrary histories.*

Proof. We only prove some cases, proofs for the others are trivial or standard.

For T_s. Take any state u with $M, u \models \overline{h}$. Then $u \in \mathcal{V}(\overline{h})$. As the sight function is reflexive, i.e., $h \in s(h)$, it holds that $\mathcal{V}(\overline{h}) \subseteq \mathcal{V}(\overline{s(h)})$. So $u \in \mathcal{V}(\overline{s(h)})$. Thus, $M, u \models \overline{s(h)}$.

For TM. Take any state u, any history h and any $z \in Z$, and suppose $M, u \models A(\overline{z} \rightarrow \overline{h})$. Then for any u', $u' \in \mathcal{V}(\overline{z})$ implies that $u' \in \mathcal{V}(\overline{h})$. Thus, $\mathcal{V}(\overline{z}) \subseteq \mathcal{V}(\overline{h})$. It follows that $z \in \mathcal{V}(\overline{h})$. Given that z is terminal, by the definition of $\mathcal{V}(\overline{h})$, it must be that $h = z$. Thus, $M, u \models A(\overline{h} \rightarrow \overline{z})$.

For DC. Take any state u, suppose for some $h_1 \lhd h_2 \lhd h_3$, $M, u \models A(\overline{h_3} \rightarrow \overline{s(h_1)})$. Then we know $\mathcal{V}(\overline{h_3}) \subseteq \mathcal{V}(\overline{s(h_1)})$. It follows that $h_3 \in s(h_1)$. As the sight function is downward closed, we have $h_2 \in s(h_1)$. Thus, $M, u \models A(\overline{h_2} \rightarrow \overline{s(h_1)})$.

For $!_{ATOM \setminus SP}$. Take any state u, and let $M, u \models [!\varphi]p$ where φ is not of the form $!\overline{s(h)}$ and p is not of the form $\overline{h_1 \geq h_2}$. It holds that $M, u \models \varphi$ implies that $M_{!\varphi}, u \models p$. By Definition 3.3, $M_{!\varphi}, u \models p$ iff $M, u \models p$. Therefore, $M, u \models \varphi$ implies $M, u \models p$. Equivalently, then, $M, u \models \varphi \rightarrow p$. □

Taut	all propositional tautologies
$T_{\geq}$	$\overline{h \geq h}$
$4_{\geq}$	$\overline{h_1 \geq h_2} \wedge \overline{h_2 \geq h_3} \to \overline{h_1 \geq h_3}$
$to_{\geq}$	$\overline{h_1 \geq h_2} \vee \overline{h_1 \geq h_2}$
T_s	$\overline{h} \to \overline{s(h)}$
TM	$\bigwedge_{z \in Z} \bigwedge_{h} (A(\overline{z} \to \overline{h}) \to A(\overline{h} \to \overline{z}))$
DC	$\bigwedge_{h_3} \bigwedge_{h_2 \lhd h_3} \bigwedge_{h_1 \lhd h_2} (A(\overline{h_3} \to \overline{s(h_1)}) \to A(\overline{h_2} \to \overline{s(h_1)}))$
NF	$\bigwedge_{h_3} \bigwedge_{h_2 \lhd h_3} \bigwedge_{h_1 \lhd h_2} (A(\overline{h_3} \to \overline{s(h_1)}) \to A(\overline{h_3} \to \overline{s(h_2)}))$
$!_{ATOM \setminus SP}$	$[!\varphi]p \leftrightarrow (\varphi \to p)$ (excluding the schema !*Sight-Preference*)
$!NEG$	$[!\varphi]\neg\psi \leftrightarrow (\varphi \to \neg[\varphi]\psi)$
$!CON$	$[!\varphi](\psi \wedge \chi) \leftrightarrow ([!\varphi]\psi \wedge [!\varphi]\chi)$
$!COM$	$[!\varphi][!\psi]\chi \leftrightarrow ![\varphi \wedge [!\varphi]\psi]\chi$
$Dual$	$[!\varphi]\psi \leftrightarrow \neg\langle !\varphi \rangle \neg\psi$

Table 1: Valid principles of L

Interpretation of valid principles. Each of these axioms has some intuitive appeal. $T_{\geq}$, 4 and $to_{\geq}$ show the *reflexivity*, *transitivity* and *totality* of the preference relation, respectively. Likewise, T_s says that sight is *reflexive*. DC characterizes the (*downward-closure*) property of sight. NF encodes the *non-forgetting* property of sight. TM guarantees that terminal histories of the P-S tree are actually *terminal*. One further interesting point is that there is no correspondence of TM for terminal histories of visible trees.

Fact 3.7. *The following formula is not valid in preference-sight models:* $\bigwedge_u \bigwedge_{z \in Z_u} \bigwedge_h (A(\overline{z} \to \overline{h}) \to A(\overline{h} \to \overline{z}))$.

Other validities in the table are axioms for standard PAL. We postpone the study of a complete axiomatization of the logic L until future work.

To conclude this section, in $\mathcal{L}$, the ingredients including histories, preferences and sights are encoded as primitive propositions. Various earlier phenomena in P-S trees can thus be captured in a simple, direct and intuitive manner. This special-purpose logic, as we will see soon, is model-dependent, but it can also be formulated generically.

4. BACKGROUND IN GAME LOGICS

In this section, we relate our logic L to existing logics for classical game theory, showing how ideas can be combined where useful. Since so far we have been working with BI and SCBI histories, we first define strategies for P-S trees:

A *strategy* for a P-S tree (T, s) is a function $\sigma : H \to A$ such that $\sigma(h) \in A(h)$. That is, σ assigns each history h an action that follows h. In particular, for a visible tree T_h, a 'local strategy' σ_h is a restriction of σ to T_h, such that $\sigma_h(h') = \sigma(h')$ for any $h' \in T_h$.

4.1 Generic formulation of $\mathcal{L}$

In applied logic for structure analysis, there exist two extremes, viz. model-dependent 'local languages' and 'generic languages' that work across models. For a generic logic, a definition of a property π is a formula φ such that for all models M, M has property π iff $M \models \varphi$. For a local language, such a formula can depend on a given model M: there exists a formula φ_M which depends on M, such that any model M has the property π iff $M \models \varphi_M$. However, in this case, the defining formula can be trivial. For example, one might define φ_M simply as follows.

$$\varphi_M = \begin{cases} \top, & \textit{if } M \textit{ satisfies } \pi \\ \bot, & \textit{Otherwise} \end{cases}$$

In this subsection, using a well-known *Rationality* property as an example, we discuss how model-dependent our earlier language L is, and then show how it can be formulated in a generic way. We first recall the results on classical BI. Given that we have been dealing with single-agent cases until now, in this Section, we will adapt the results from the literature on multi-player games to the single-player case.

The BI strategy [22; 23] is the largest subrelation σ of the total *move* relation that has at least one successor at each node, while satisfying the rationality (RAT) property:

RAT No alternative move for the player yields an outcome via further play with σ that is strictly better than all the outcomes resulting from starting at the current move and then playing σ all the way down the tree.

As argued in [22; 23], this rationality assumption is a confluence property for action and preference:

$$\textbf{CF} \qquad \forall x \forall y (x \sigma y \to \forall z (x \textit{ move } z \to$$
$$\exists u (end(u) \wedge y \sigma^* u \wedge \forall v ((end(v) \wedge z \sigma^* v) \to u \geq v)))).$$

We can observe that there is also a corresponding rationality property for the local BI strategies that constitute an SCBI, which should however now express a confluence property for action, preference and sight. Specifically, for a P-S tree, each local BI strategy for the visible tree T_h at h is the largest subrelation σ_h of the total *move* relation in T_h, satisfying 1) σ_h has at least one successor at each $h' \in T_h$, and 2) the following rationality property holds:

RATS In the visible tree, there is one outcome obtained by playing σ_h from the start to the end, that is no worse than all the outcomes yielded from any alternative first move followed by further play with σ_h.

This confluence property involving sight is expressible as follows in our language $\mathcal{L}$:

Proposition 4.1. *Let (T, s) be a P-S tree, and let M be any model for it. M satisfies* RATS *iff M validates the following $\mathcal{L}$-formula, where σ_h is the* BI *strategy for visible tree at h and where $(h(\sigma_h)^k)$ stands for the history reached from h after executing σ_h for k times.*

$$\textbf{CFS}_M \quad \bigwedge_{h} \bigvee_{z \in Z_h} \bigvee_{k = l(z) - l(h)} (A(\overline{(h(\sigma_h)^k)} \leftrightarrow \overline{z})$$
$$\to (\bigwedge_{a' \in A_h(h)} \bigwedge_{z' \in Z_h} \bigwedge_{m = l(z') - l(ha')} (A(\overline{(ha'(\sigma_h)^m)} \leftrightarrow \overline{z'})) \to$$
$$\overline{z \geq z'})).$$

Proof. We first claim that in any preference-sight model M, and state $h \in H$, for any terminal history $z \in Z_h$, and $h' \in H_h$, $A(\overline{h'} \leftrightarrow \overline{z})$ implies that $h' = z$. This is straightforward since $A(\overline{h'} \leftrightarrow \overline{z})$ demonstrates that prefixes of h' are the same with those of z, which means that $h' = z$. Then $M \models \textbf{CFS}_M$ says that there is a terminal history z_h following h by playing a local BI strategy σ_h, such that $z \succeq z'$ for any other $z' \in Z_h$ which follows an alternative first move $a' \in A_h(h)$ via further play of σ_h. Therefore, we know that M satisfies **RATS**. □

However, compared with the generic logic in [24; 23; 22], the given definition in our logic is local. It is obvious that **CF**, the formula defining the property **RAT**, is insensitive to models – while our $\mathbf{CFS}_M$ relies on a given model for its ranges of big disjunctions and conjunctions, and in its model-dependent notations like $s(h)$ and $\overline{h_1 \geq h_2}$. Still, it is also clearly true that our definition is not as trivial as the earlier local trick. Therefore, our logic $\mathcal{L}$ seems somewhere between the two extremes of locality and genericity. This feeling can be made precise by moving to a closely related truly generic first-order logic of preference-sight trees.

The relevant modified formula involves some natural auxiliary predicates. $x \lhd y$ says that x is a prefix of y; $x \lessdot y$ means that x can see y. Corresponding to the BI relation σ, $y\sigma(x)z$ says that from y, z is a local backward induction move in the visible tree at x; σ^k describes σ being composed for k times with $k \in \mathbb{N}$ [4]; *move* and $\geq$ are still the move relation and preference relation, respectively, of the game.

Proposition 4.2. *Any model M satisfies* **RATS** *iff it validates the following formula.*

$$\mathbf{CFS}(FO): \qquad \forall x\{(\exists y(x \lhd y)) \rightarrow$$
$$\forall u[(x\sigma(x)u) \rightarrow \forall t((x \ \mathit{move} \ t \wedge x \lessdot t) \rightarrow$$
$$\exists z((x \lessdot z \wedge \neg\exists z'(z \lhd z' \wedge x \lessdot z') \wedge \exists k(u(\sigma(x))^k z))\wedge$$
$$\forall v((x \lessdot v \wedge \neg\exists v'(v \lhd v' \wedge v \lessdot v') \wedge \exists l(t(\sigma(x))^l v)) \rightarrow$$
$$\wedge z \geq v))))]\}.$$

Proof. It is easy to show that

$$M \models \mathbf{CFS}(FO) \ \mathit{iff} \ M \models \mathbf{CFS}_M. \quad \square$$

In summary, incorporating basic elements of P-S trees directly into first-order syntax makes L intuitive and natural.

Even so, other logics exist for dealing with further aspects of game trees and solution procedures, and we will discuss a few examples in what follows with a view to how they behave in the presence of sight.

4.2 Solution procedures and fixed-point logics

Recursive solution procedures naturally correspond to definitions in existing fixed-point logics, such as the widely used system LFP(FO). An LFP(FO) formula mirroring the recursive nature of BI is constructed in [24; 26] to define the classical BI relation, based on the above property **RAT**. Now, we have shown that sight-restricted SCBI, too, is a recursive game solution procedure. Can LFP(FO) be used to define SCBI as well – and if so, how?

The answer is yes, but we need an extension. Rather than a binary relation *bi* as in [24; 26], characterizing SCBI needs a *ternary* relation. First, we define the local BI relation in visible trees, which will be denoted by bi_{sight}. For any states x, y, z, $bi_{sight}(x, y, z)$ means that in the visible tree at x, the local BI strategy is bi_{sight}, which chooses z when the current state is y. It is then obvious that bi_{sight} should satisfy the following simple first-order definable property, requiring the relevant states to be visible and reachable:

$$bi_{sight}(x, y, z) \rightarrow \mathit{see}(x, y) \wedge \mathit{see}(x, z) \wedge \mathit{move}(y, z).$$

The intuition of $bi_{sight}(x, y, z)$ is then captured as follows:

$$\forall x\forall y\forall z(bi_{sight}(x, y, z) \rightarrow \forall t((\mathit{see}(x, t) \wedge \mathit{move}(y, t))$$
$$\rightarrow (\exists u(\mathit{end}_{sight}(x, u) \wedge bi^*_{sight}(x, z, u) \wedge \forall v((\mathit{end}_{sight}(x, v)\wedge$$
$$bi^*_{sight}(x, t, v)) \rightarrow u \geq v))))).$$

Notice that all occurrences of bi_{sight} in the above formulas are still syntactically positive. This allows us to define local BI strategy bi_{sight} with LFP(FO).

Proposition 4.3. *The strategy bi_{sight} can be defined as the relation R in the following* LFP(FO) *formula.*

$$\nu R, xyz \bullet \forall x\forall y\forall z(R(x, y, z) \rightarrow \forall t((\mathit{see}(x, t) \wedge \mathit{move}(y, t))$$
$$\rightarrow (\exists u(\mathit{end}_{sight}(x, u) \wedge R^*(x, z, u) \wedge \forall v((\mathit{end}_{sight}(x, v)\wedge$$
$$R^*(x, t, v)) \rightarrow u \geq v))))).$$

It can be proved formally that bi_{sight} is a greatest-fixed-point of the above formula. Based on bi_{sight}, we now proceed to show that the SCBI relation is LFP(FO) definable.

Corollary 4.4. *The* SCBI *relation scbi for a P-S tree can be represented in the following formula:*

$$\forall x\forall y(\mathit{scbi}(x, y) \leftrightarrow bi_{\mathrm{sight}}(x, x, y)).$$

As in the original classical case, this LFP(FO) definability of *scbi* exposes an intersection between the logical foundation of computation and the recursive nature of sight-compatible backward induction solutions for P-S trees.

4.3 Modal surface logic of best action

In contrast with detailed formalism of solutions with LFP(FO), there is the modal surface logic of [25], which enables direct and natural reasoning about best actions without considering the underlying details of recursive computation. First of all, we list its modalities for classical BI. $[bi]$ and $[BI]$ encode the BI move and BI paths respectively. $[best]\varphi$ says that φ is true in some successor of the current node that can be reached in one step via the *bi* move.

$M, h \models \mathit{end}$ iff $h \in Z$.
$M, h \models [\mathit{move}]\varphi$ iff $\forall$ $h' = (ha)$ with $a \in A(h)$, $M, h' \models \varphi$.
$M, h \models [\mathit{best}]\varphi$ iff for all h' with $h' \in bi(h)$, $M, h' \models \varphi$.
$M, h \models [bi]\varphi$ iff for all h' with $h' \in bi(h)$, $M, h' \models \varphi$.
$M, h \models [bi^*]\varphi$ iff $M, u \models \varphi$ for all u with $u \in (bi)^*(h)$.
$M, h \models [BI]\varphi$ iff for all z with $z \in \mathbf{BI}$, $M, z \models \varphi$.

The above logic is still applicable in our setting, but it requires substantial extension for sight-related concepts. In accordance with $[bi]$ and $[BI]$, we use $[\mathit{scbi}]$ and $[\mathit{SCBI}]$ as operators for the SCBI strategy and SCBI path, respectively. For the local BI strategy and path in visible trees, the modalities are $[bi_{sight}]$ and $[BI_{sight}]$. Moreover, recall that $M_{!s(h)}$ is the updated model obtained in the way of Definition 3.3.

$M, h \models [\mathit{scbi}]\varphi$ iff for all h' with $h' \in \mathit{scbi}(h)$, $M, h' \models \varphi$.
$M, h \models [\mathit{SCBI}]\varphi$ iff for all h' with $z \in \mathbf{SCBI}$, $M, z \models \varphi$.
$M, h \models [!\mathit{sight}]\varphi$ iff $M_{!s(h)}, h \models \varphi$.
$M_{!s(h)}, u \models \mathit{end}_{sight}$ iff $u \in Z_h$.
$M_{!s(h)}, u \models [\mathit{move}_{sight}]\varphi$ iff for $\forall u' = (ua)$ *with* $a \in A_h(u)$,
$M_{!s(h)}, u' \models \varphi$.
$M_{!s(h)}, u \models [\mathit{best}_{sight}]\varphi$ iff $M, u' \models \varphi$ for $\forall u' \in bi_h(u)$.
$M_{!s(h)}, u \models [bi_{sight}]\varphi$ iff $M, u' \models \varphi$ for $\forall u' \in bi_h(u)$.
$M_{!s(h)}, u \models [(bi_{sight})^*]\varphi$ iff $M_{!s(h)}, u' \models \varphi$ for all u',

[4] Here $x\sigma^k y$ is the abbreviation of $\exists y_1\exists y_2 \cdots \exists y_k(x\sigma y_1 \wedge y_1\sigma y_2 \wedge \cdots \wedge y_{k-1}\sigma y_k \wedge (y_k = y))$.

such that $u' \in (bi_h)^*(u)$.

$M, h \models [BI_{sight}]\varphi$ iff for all z with $z \in \mathbf{BI}_h$, $M, z \models \varphi$.

We give a few illustrations of new issues that arise now.

Capturing the SCBI strategy For a start, we are now able to characterize the SCBI strategy, in a similar vein as the frame correspondence for the classical BI strategy in [25].

Proposition 4.5. *The* BI *strategy is the unique relation bi satisfying this modal axiom for all propositions p:*

$$(\langle bi^* \rangle (end \wedge p)) \rightarrow ([move][\sigma^*](end \wedge \langle \leq \rangle p)).$$

Along the same lines, we can express the SCBI strategy in P-S trees based on the idea that each *scbi* move coincides with a local BI move within the current visible tree.

Proposition 4.6. *The* SCBI *strategy is the relation scbi satisfying the following axioms for all propositions p:*

(1) $$\langle scbi \rangle p \leftrightarrow [!sight]\langle bi_{sight} \rangle p.$$

(2) $$[!sight](\langle (bi_{sight})^* \rangle (end_{sight} \wedge p) \rightarrow [move_{sight}]\langle (bi_{sight})^* \rangle (end_{sight} \wedge \langle \leq \rangle p)).$$

Best action and preference-consistency Turning to properties of frames for the extended modal logic of best action with sight, there are interesting differences when comparing SCBI and classical BI. To see this, we employ operators $\langle best \rangle$, $\langle bi^* \rangle$, $\langle scbi^* \rangle$, $\langle best_{sight} \rangle$ and $\langle (bi_{sight})^* \rangle$. Now we can make some interesting comparisons.

Proposition 4.7. *For classical backward induction, the axiom $\langle best \rangle \langle bi^* \rangle \varphi \leftrightarrow \langle bi^* \rangle \varphi$ holds.*

However, the new frames do not have the corresponding axiom for the SCBI strategy, since the actions it recommends are not necessarily the actual best actions according to BI. Even in visible trees, this is also not true.

Proposition 4.8. *The following formulas are not valid:*

$$(a) \quad \langle best \rangle \langle scbi^* \rangle \varphi \leftrightarrow \langle scbi^* \rangle \varphi.$$

$$(b) \quad [!sight](\langle best \rangle \langle (bi_{sight})^* \rangle \varphi \leftrightarrow \langle (bi_{sight})^* \rangle \varphi).$$

Nevertheless, there is a certain coherence between the local BI strategy and local best actions returned by it.

Proposition 4.9. *The following formula is valid:*

$$[!sight](\langle best_{sight} \rangle \langle (bi_{sight})^* \rangle \varphi \leftrightarrow \langle (bi_{sight})^* \rangle \varphi).$$

As for the preference relation, SCBI has a property that classical BI lacks: local BI moves never conflict with the preferences in submodels. In other words, within a visible tree, the initial move determined by the local BI strategy is more preferable for the agent than any other first move.

Proposition 4.10. *For* SCBI*, it holds that*

$$[!sight](\langle best_{sight} \rangle \varphi \rightarrow [move_{sight}]\langle \leq \rangle \varphi).$$

For BI, although it returns a final optimal path, there is no guarantee that its intermediate histories be preferable.

Proposition 4.11. *For* BI*, the following does not hold:*

$$\langle best \rangle \varphi \rightarrow [move]\langle \leq \rangle \varphi.$$

Path terminality and optimality Using a similar style of modal analysis, we can make the following observations concerning the obvious operators $[BI]$, $[SCBI]$ and $[BI]_{sight}$.

Proposition 4.12. *We have the following three facts:*

(a) *The formula$[BI]\varphi \rightarrow [BI][BI]\varphi$ is valid.*

(b) *For* SCBI*, the following formula does not hold:*

$$[BI_{sight}]\varphi \rightarrow [BI_{sight}][BI_{sight}]\varphi.$$

(c) *The formula $[SCBI]\varphi \rightarrow [SCBI][SCBI]\varphi$ is valid.*

Here (a) says that from a BI outcome only a terminal history can be reached; (b) shows that the local BI history may not be a terminal history of the whole tree, and (c) says the SCBI history for the whole tree is always terminal.

Another phenomenon regarding these operators is the local optimality of SCBI at the cost of being more realistic than BI. We have mentioned this point already in Section 2.2.4: now we can present a precise formal version.

Proposition 4.13. *Let σ be any strategy profile,*

(a). *For* BI*, the following is valid:*

$$\langle BI \rangle \varphi \rightarrow [\sigma]\langle \leq \rangle \varphi.$$

(b). *The following does not hold:*

$$\langle SCBI \rangle \varphi \rightarrow [\sigma]\langle \leq \rangle \varphi.$$

(c). *For* SCBI*, it holds that*

$$[!sight](\langle BI_{sight} \rangle \varphi \rightarrow [\sigma_{sight}]\langle \leq_{sight} \rangle \varphi).$$

Here (a) shows the global optimality of the BI path. (b) and (c) together say the SCBI path is not globally optimal, but each move on this path leads to a locally optimal path.

Altogether, this section has shown the broad logical foundations of our framework, embedding our local language in existing broader generic formalisms, but also enriching and extending these frameworks with aspects of short sight.

5. TOWARD MULTI-PLAYER GAMES

While our models and results are about single-agent sequential decision-making processes, we believe they are applicable well beyond that. They can be naturally extended to multi-player extensive game-scenarios with short sight. For such a game model, we can build on [6], which makes an assumption that the current player only knows his own sight, and that he believes other players can see as much as he can see and will play according to this belief. That is, this model precludes more complex forms of interactive knowledge and reasoning. But using this same assumption, our model in this paper can be extended to multi-player cases directly. The only thing we have to do is add agent-labeling to SSDM: even though players can change with time, everything including sight, preference, and actions can be modeled from the current player's perspective.

We will not state any results for the extended multi-player model since they are quite similar to what we have shown already. The case where we drop the above assumption and allow a more free modeling of players' mutual knowledge and beliefs about sight and preference would be more interesting. We will leave this for future work.

6. DISCUSSION AND CONCLUSION

Though motivated by single-agent decision-making process, we have gone towards a much more general goal. In the process, our analysis significantly adds to current connections between logic, computation, and game solutions.

In many recent game-theoretic papers centering on *bounded rationality*, a model has been used of *games with awareness*, [7; 12; 5; 8]. This approach generalizes the classical representation of extensive games by modeling players who may not be aware of all the paths. While [6] shows that games with short sight are a well-behaved subclass of games with awareness, there exists a fundamental difference in focus. Players in the latter approach may be unaware of some branches but they can always see some terminal histories, while in the former, players' sight may only include intermediate histories, ruling out all terminal ones. Moreover, we have shown how short-sight games allow for a natural co-existence of two views of a game, that of insiders and that of outsiders. Having said this, it is clearly an interesting issue to see if our approach in this paper can be extended to cover awareness.

Another obvious interface for our logics are heuristic evaluation approaches for intermediate nodes used by the AI community for computational game-solving, [15; 21]. This, too, is a connection that deserves further exploration.

There are many additional topics to pursue. For instance, we already mentioned multi-player scenarios with non-trivial interactive reasoning about other agents' preferences, sights, and strategies. This has also been identified as a key task for epistemic game theory.

Acknowledgments

I thank Fenrong Liu for our fruitful collaboration on earlier versions of this paper. Paolo Turrini provided crucial insights on short-sight games and their connections with games and computation, which we are partly exploring together. Sonja Smets provided helpful comments overall. But especially, I thank Johan van Benthem for our longstanding contacts on the logic of short-sight games: Section 4 of this paper owes a lot to his many suggestions and observations. This work is supported by the China Scholarship Council and NSFC grant No. 61472369.

7. REFERENCES

[1] A. Baltag, S. Smets, and J. A. Zvesper. Keep 'hoping' for rationality: a solution to the backward induction paradox. *Synthese*, 169(2):301–333, 2009.

[2] J. L. Bermúdez. *Decision Theory and Rationality.* Oxford University Press, 2009.

[3] P. Blackburn, M. de Rijke, and Y. Venema. *Modal logic.* Cambridge University Press, 2001.

[4] C. Degremont, S. Paul, and N. Asher. A Logic of Sights. *Journal of Logic and Computation*, 2014.

[5] Y. Feinberg. Games with unawareness. *Stanford Graduate School of Busirness Paper No. 2122*, 2012.

[6] D. Grossi and P. Turrini. Short sight in extensive games. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, pages 805–812, 2012.

[7] J. Y. Halpern and L. C. Rêgo. Extensive games with possibly unaware players. In *AAMAS*, pages 744–751, 2006.

[8] J. Y. Halpern and L. C. Rêgo. Extensive games with possibly unaware players. *Mathematical Social Sciences*, 70:42–58, 2014.

[9] S. O. Hansson. Decision theory -a brief introduction, 1994.

[10] P. Harrenstein, W. V. D. Hoek, J. jules Meyer, and C. Witteveen. On modal logic interpretations of games. In *Procs ECAI 2002*, pages 28–32, 2002.

[11] P. Harrenstein, W. van der Hoek, J.-J. Meyer, and C. Witteveen. A modal characterization of nash equilibrium. *Fundam. Inf.*, 57(2-4):281–321, 2003.

[12] A. Heifetz, M. Meier, and B. C. Schipper. Dynamic unawareness and rationalizable behavior. *Games and Economic Behavior*, 81:50–68, 2013.

[13] B. Houlding. *Sequential Decision Making with Adaptive Utility.* PhD thesis, Department of Mathematical Sciences, Durham University, 2008.

[14] K. Høyland and S. W. Wallace. Generating scenario trees for multistage decision problems. *Management Science*, 47(2):pp. 295–307, 2001.

[15] Y. J. Lim and W. S. Lee. Properties of forward pruning in game-tree search. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, AAAI'06, pages 1020–1025. AAAI Press, 2006.

[16] M. L. Littman. *Algorithms for Sequential Decision-making.* PhD thesis, Brown University, Providence, RI, USA, 1996.

[17] C. Liu, F. Liu, and K. Su. A logic for extensive games with short sight. In *LORI*, pages 332–336, 2013.

[18] D. W. North. A tutorial introduction to decision theory. *IEEE Transactions on Systems Science and Cybernetics*, 1968.

[19] M. J. Osborne. *An Introduction to Game Theory*, volume 2. Oxford University Press, 2004.

[20] M. J. Osborne and A. Rubinstein. *A Course in Game Theory.* MIT Press, 1994.

[21] F. Rossi, K. B. Venable, and T. Walsh. *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice.* Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011.

[22] J. van Benthem. Exploring a theory of play. In *Proc. of TARK*, pages 12–16, 2011.

[23] J. van Benthem. *Logic in Games.* MIT Press, 2014.

[24] J. van Benthem and A. Gheerbrant. Game solution, epistemic dynamics and fixed-point logics. *Fundam. Inform.*, 100(1-4):19–41, 2010.

[25] J. van Benthem, S. V. Otterloo, and O. Roy. Preference logic, conditionals, and solution concepts in games. In *Modality Matters*, pages 61–76. University of Uppsala, 2006.

[26] J. van Benthem, E. Pacuit, and O. Roy. Toward a theory of play: A logical perspective on games and interaction. *Games*, 2(1):52–86, 2011.

[27] W. van der Hoek and M. Pauly. Modal logic for games and information. In J. van Benthem, P. Blackburn, and F. Wolter, editors, *Handbook of Modal Logic.* Elsevier, 2006.

[28] H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese library.* Springer, 2007.

The optimality of coarse categories in decision-making and information storage

Michael Mandler
Department of Economics
Royal Holloway College, University of London
Egham, Surrey, United Kingdom

1. INTRODUCTION

Suppose that agents, rather than forming a separate preference judgment for each pair of alternatives, make decisions using *criteria*. A criterion orders a small number of categories, each of which consists of many alternatives. The potential of a criterion to order alternatives within another criterion's categories allow sets of criteria to generate large numbers of choice distinctions. If an agent has objective preferences that can be inferred from a large set of sufficiently discriminating criteria, the agent will be better off if more of the criterion orderings are discovered: the agent will then be able to determine the optimal allocation from more choice sets. The uncovering of more criterion discriminations is costly, however, and we therefore consider *efficient* points where the cost of making a given number of choice distinctions is minimized.

This optimization problem seems to lead to a trade-off. Given a number of choice distinctions, an agent could either use a large set of coarse criteria (criteria with only a small number of categories) or a small set of finer, more discriminating criteria. We show under mild conditions that large sets of coarse criteria always lead to reductions in decision-making costs. Binary criteria with only two categories per criterion therefore provide the only efficient arrangement. Under mild restrictions on how criteria are aggregated into decisions, binary criteria lead to *rational* choice functions, where decisions are determined by a complete and transitive binary relation.

We apply our model to the problem of determining the optimal number of digits in an information storage device. We show that, even if the marginal cost of additional digits declines rapidly to 0, binary digits (bits) offer the efficient solution.

In this short paper, we pay particular attention to the symmetry conditions that are entailed when sets of criteria are efficient. A full working paper [2] is available on-line.

2. AN OUTLINE OF THE MODEL

A *criterion* C_i is an asymmetric binary relation on a domain of alternatives X and a *set of criteria* is denoted $\mathcal{C} = \{C_1, ..., C_N\}$. Two alternatives x and y are deemed C_i-equivalent if x and y share the same set of C_i-superior alternatives and the same set of C_i-inferior alternatives (see [1]). A *C_i-category* is a maximal set of C_i-equivalent alternatives and $e(C_i)$ denotes the number of C_i-categories. The *discrimination vector* of $\mathcal{C}$ is $(e(C_1), ..., e(C_N))$. A C_i is *coarser* than C'_i if $e(C_i) < e(C'_i)$.

Let c be a choice function on a domain of finite subsets of X. Two alternatives x and y are in the same *choice class* of c if c treats them interchangeably: first, when x is chosen and y is available then y is chosen too, and second, if x but not y is available then x is chosen if and only if, when y is available and not x, y is chosen.

A choice function c *uses* $\mathcal{C}$, denoted $(\mathcal{C}, c)$, if c does not make distinctions that are not already present in the criteria: for each set of alternatives A that contains only alternatives that are in the same C_i-category, $i = 1, ..., N$, there is a choice class of c that contains A.

Let $\kappa(C_i)$ denote the *cost of criterion* C_i. We assume $\kappa(C_i)$ is determined by the number of C_i-categories and therefore also write $\kappa(e)$ to denote the cost of a C_i with e categories. We assume that the *cost of a set of criteria*, $\kappa[\mathcal{C}]$, equals the sum of the costs of the criteria in $\mathcal{C}$. Letting $n(c)$ be the number of choice classes in c, a pair $(\mathcal{C}, c)$ is *more efficient* than the pair $(\mathcal{C}', c')$ if $n(c) \geq n(c')$ and $\kappa[\mathcal{C}] \leq \kappa[\mathcal{C}']$, with at least one strict inequality, and $(\mathcal{C}, c)$ is *efficient* if there does not exist a $(\mathcal{C}', c')$ that is more efficient than $(\mathcal{C}, c)$.

The fundamental advantage of criteria is that each criterion can discriminate within the categories of other criteria. Given constraints that specify that criterion C_i can have no more than e_i categories (and assuming that $|X|$ is sufficiently large), we can find a $(\mathcal{C}, c)$ such that (i) there is a partition of X with $\prod_{i=1}^{N} e_i$ cells such that x and y are in distinct cells if and only if they lie in different C_i-categories for at least one i and (ii) each cell of this partition forms a choice class of c. Subject to the e_i constraints, this $(\mathcal{C}, c)$ maximizes $n(c)$ and accordingly we define $(\mathcal{C}, c)$ to *maximally discriminate* if $n(c) = \min \left[\prod_{i=1}^{N} e(C_i), |X|\right]$.

3. MAIN RESULTS

(1) Since criteria with only one category make no discriminations and require no decisions, we assume they are costless. To compare a $(\mathcal{C}, c)$ and $(\mathcal{C}', c')$ that have the same number of costly categories, suppose that $\sum_{i=1}^{N} (e(C_i) - 1) = \sum_{i=1}^{N'} (e(C'_i) - 1)$. Assume also that either (i) the marginal cost of categories is increasing and the smaller of $n(c)$ and $n(c')$ is less than the cardinality of X or (ii) the marginal costs of categories is strictly increasing. We show that if $\mathcal{C}$ has greater proportions of coarser criteria than does $\mathcal{C}'$ and if $(\mathcal{C}, c)$ maximally discriminates, then $(\mathcal{C}, c)$ is more efficient than $(\mathcal{C}', c')$.

(2) Fix a set of domains that, for each finite m, contains a X with m elements and call a domain *admissible* if it is drawn from this set. Then, every efficient $(\mathcal{C}, c)$ where

the domain is admissible has a $\mathcal{C}$ that contains only binary criteria if and only if $\kappa(e) > \kappa(2)\lceil\log_2 e\rceil$ for all integers $e > 2$.

Thus the cost of e categories can rise as slowly as $\log_2 e$ – in which case the marginal cost of categories descends to 0 – and still the only efficient arrangement is for all criteria to be binary.

(3) We apply the result in (2) to information storage. Suppose we wish to store some integer between 1 and n using N k-ary digits and that the cost of storage equals $\kappa(k)N$. We show that, for all positive integers n, binary digits are the minimum-cost storage method if and only if $\kappa(k) > \kappa(2)\lceil\log_2 k\rceil$ for all integers $k > 2$.

(4) We specify axioms for how to aggregate sets of criteria into choice functions that generalize weighted voting. Suppose that the choice function c in the pair $(\mathcal{C}, c)$ satisfies these axioms, that $\mathcal{C}$ contains only binary criteria, and that c satisfies the following Condorcet rule: if there is a x in a choice set A that is chosen by c from all $\{x, y\}$ with $y \in A$ then x is chosen from A too. Then c makes selections that maximize a complete and transitive binary relation. Given (2), we conclude that in a broad range of cases, efficient decision-making is rational.

4. SYMMETRY AND MAXIMAL CATEGORIZATION

Maximal discrimination is necessary for decision-making efficiency since otherwise $n(c)$ could be increased without an increase in costs. The key feature required for a $(\mathcal{C}, c)$ to maximally discriminate is that the following property of $\mathcal{C}$, called *maximal categorization*, is satisfied: the *discrimination partition* $\mathcal{P}$ of X that places x and y in distinct cells if and only if x and y lie in different C_i-categories for at least one i must have $\prod_{i=1}^{N} e(C_i)$ cells.

We will now see that if X is a product of attributes and each criterion orders a distinct attribute, then maximal categorization is satisfied and conversely if maximal categorization is satisfied then we can label alternatives so that X becomes a product of attributes. By joining this conclusion to result (2), that only binary criteria are efficient, we can describe efficient decision-making concisely: to be efficient agents must be able to describe the alternatives in X so that they form a product of attributes and each criterion must divide a distinct attribute into exactly two categories.

The simplest way to achieve maximal categorization is for X to be formed by a product of attributes and for each C_i to divide X into categories based only on attribute i. The domain X might be a set of cars, and the attributes might be colors, top speeds, and prices. A 'speed' C_i would then order cars based on the ranges of top speeds that C_i deems to be equivalent.

Formally, an *attribute* is a set X_i and N attributes define the domain of alternatives $X = \prod_{i=1}^{N} X_i$. We will say that a set of criteria $\mathcal{C}$ *is based on a product of attributes* if for each C_i there is a set X_i and a partition of $\{X_i^1, ..., X_i^{e(C_i)}\}$ of X_i such that the categories of C_i are the sets $X_i^j \times \left(\prod_{k \neq i} X_k\right)$, $j = 1, ..., e(C_i)$. So, if C_i is an ordering of cars by color then each X_i^j would represent a color and x and y would be placed into distinct C_i-categories if and only if the ith coordinates of x and y indicate different colors: $x_i \in X_i^j$ and $y_i \in X_i^k$ where $j \neq k$. The cells of the discrimination partition $\mathcal{P}$ would then be the $\prod_{i=1}^{N} e(C_i)$ sets $X_1^{j_1} \times \cdots \times X_N^{j_N}$ where, for each i, j_i is an integer between 1 and $e(C_i)$. Maximal categorization thus obtains.

This treatment assumes that X is a product space: for each possible combination of attributes (each possible color-speed-price combination), there is a corresponding element of X. But for maximal categorization it is enough that there is *some* alternative in X for each combination of attribute ranges specified by the criteria, that is, it is sufficient for X to be a subset of $\prod_{i=1}^{N} X_i$ such that each $X_1^{j_1} \times \cdots \times X_N^{j_N}$ intersects X.

A set of criteria $\mathcal{C}$ that is based on a product of attributes enjoys a wide-ranging symmetry property. Fix some C_i in $\mathcal{C}$, and consider a set $\mathcal{E}_{-i}$ formed by an arbitrary union of the categories of the remaining criteria C_j, $j \neq i$. Given the product structure of $\mathcal{C}$, any such $\mathcal{E}_{-i}$ must intersect each of the C_i-categories. To continue the car example, the set of cars $\mathcal{E}_{-i}$ defined by a certain range of top speeds and prices can be partitioned into all the possible color subsets, say blue, red, and yellow. If we use C_i to order the cells of the color partition of the cars in $\mathcal{E}_{-i}$, the ordering will have the same 'shape' as – be order isomorphic to – the original color ordering C_i of X. If, for example, C_i on X is a cycle – blue is better than red which is better than yellow which is better than blue – then the C_i ordering of any set of cars defined by a range of speeds and prices will also form a cycle. We conclude that any two sets of cars Y and Z defined by selections of non-color attributes will be order isomorphic to each other when each is endowed with the color ordering C_i (or rather the restrictions of C_i to Y and Z).

This symmetry property may seem to be of limited value since it appears to apply only to products of attributes. But in fact the symmetry property characterizes any $\mathcal{C}$ that maximally categorizes. If for an arbitrary (possibly nonproduct) $\mathcal{C}$, we apply C_i to some $\mathcal{E}_{-i}$ and it defines fewer than $e(C_i)$ C_i-category subsets then the discrimination partition $\mathcal{P}$ would have to contain fewer than $\prod_{i=1}^{N} e(C_i)$ cells. And the only way that $\mathcal{E}_{-i}$ and $\mathcal{E}'_{-i}$ can each define $e(C_i)$ C_i-category subsets is for the C_i ordering of these subsets to be order isomorphic.

Moreover, if an arbitrary (possibly nonproduct) $\mathcal{C}$ enjoys the symmetry property we can relabel the elements of the domain X so that $\mathcal{C}$ is then based on a product of attributes. To do this, we associate each C_i with an attribute (e.g., color) and identify each C_i-category with an arbitrary value X_i^j for that attribute (e.g., blue): each cell of $\mathcal{P}$ is thus identified with a vector of attribute values. So, although a product of attributes looks special, it provides a model for any set of criteria that maximally categorizes.

The following definitions and theorem make these claims precise. We use $E_i^1, ..., E_i^{e(C_i)}$ to denote the categories of criterion C_i.

Given the set of criteria $\{C_1, ..., C_N\}$, $\mathcal{E}_{-i}$ is a *union of C_{-i}-categories* if $\mathcal{E}_{-i} = \bigcup_j E_k^j$ for some collection of criterion categories $\{E_k^j\}$ such that $k \neq i$ for each j. Let $C_i^{\mathcal{E}_{-i}}$ denote the binary relation defined by $E\, C_i^{\mathcal{E}_{-i}} E'$ if and only if there are C_i-categories E_i and E'_i such that $E = E_i \cap \mathcal{E}_{-i}$, $E' = E'_i \cap \mathcal{E}_{-i}$, and $x\, C_i\, y$ for $x \in E$ and $y \in E'$. We then say $\mathcal{C}$ satisfies the *order-isomorphism property* if for any i and any two unions of C_{-i}-categories, $\mathcal{E}_{-i}$ and $\mathcal{E}'_{-i}$, the binary relations $C_i^{\mathcal{E}_{-i}}$ and $C_i^{\mathcal{E}'_{-i}}$ are order-isomorphic.

The set of criteria $\mathcal{C}$ has a *product representation* if (i) for each i, there is a nonempty set Y_i and a partition $\{Y_i^1, ...,$

$Y_i^{e(C_i)}\}$ of Y_i, (ii) there is a set of criteria $\widehat{\mathcal{C}} = \{\widehat{C}_1, ..., \widehat{C}_N\}$ defined on $Y = \prod_{i=1}^{N} Y_i$ where the categories of each $\widehat{C}_i$ are the sets $Y_i^j \times \left(\prod_{k \neq i} Y_k\right)$, and (iii) for each i, there is a order-preserving bijection f between the categories of C_i and $\widehat{C}_i$, that is, $E_i^j C_i E_i^{j'}$ if and only if $f\left(E_i^j\right) \widehat{C}_i f\left(E_i^{j'}\right)$.

We then have the following result.

Theorem. For a set of criteria $\mathcal{C}$, the following statements are equivalent: (i) $\mathcal{C}$ maximally categorizes, (ii) $\mathcal{C}$ satisfies the order-isomorphism property, (iii) $\mathcal{C}$ has a product representation.

5. REFERENCES

[1] Mandler, M., Rational agents are the quickest. J. Econ. Theory 155: 206-233, 2015.

[2] Mandler, M., Coarse, efficient decision-making. Available at SSRN: http://ssrn.com/abstract=2494600 or http://dx.doi.org/10.2139/ssrn.2494600.

Undecidable Cases of Model Checking Probabilistic Temporal-Epistemic Logic

Ron van der Meyden
School of Computer Science and Engineering
UNSW Australia

Manas K Patra
School of Computer Science and Engineering
UNSW Australia

ABSTRACT

We investigate the decidability of model-checking logics of time, knowledge and probability, with respect to two epistemic semantics: the clock and synchronous perfect recall semantics in partially observed discrete-time Markov chains. Decidability results are known for certain restricted logics with respect to these semantics, subject to a variety of restrictions that are either unexplained or involve a longstanding unsolved mathematical problem. We show that mild generalizations of the known decidable cases suffice to render the model checking problem definitively undecidable. In particular, for a synchronous perfect recall, a generalization from temporal operators with finite reach to operators with infinite reach renders model checking undecidable. The case of the clock semantics is closely related to a monadic second order logic of time and probability that is known to be decidable, except on a set of measure zero. We show that two distinct extensions of this logic make model checking undecidable. One of these involves polynomial combinations of probability terms, the other involves monadic second order quantification into the scope of probability operators. These results explain some of the restrictions in previous work.

1. INTRODUCTION

Model checking is a verification methodology used in computer science, in which we ask whether a given model satisfies a given formula of some logic. First proposed in the 1980's [CE81], model checking is now a rich area, with a large body of associated theory and well developed implementations that automate the task of model checking. Significant use of model checking tools is made in industry, in particular, in the verification of computer hardware designs.

Model checking developed originally in a setting where the specifications are expressed in a propositional temporal logic, and the systems to be verified are finite state automata. This setting has the advantage of being decidable, and a great deal of work has gone into the development of algorithms and heuristics for its efficient implementation. More recently, the field has explored the extent to which the expressiveness of both the model representations and of the specification language can be extended while retaining decidability of model checking. Extensions in the systems dimensions considered include real-time systems [ACD90], systems with a mixed continuous and discrete dynamic [MNP08], richer automaton models such as push-down automata, machines with first-in-first out queues etc. In the dimension of the specification language, extensions considered include elements of second order logic and specific constructs to capture the richer properties of the systems models described above (e.g. in the real time case the specification language might contain inequalities over time values.)

Model checking for epistemic logic was first mooted in [HV91], and model checking for the combination of temporal and epistemic logic has been developed both theoretically [MS99, EGM07, HM10] and in practice [GM04, LQR09, KNN$^+$08, Eij04]. A variety of semantics for knowledge are known to be associated with decidable model checking problems in finite state systems, in particular, the observational semantics (in which an agent reasons based on its present observation) the clock semantics (in which an agent reasons based on its present observation and the present clock value), and synchronous and asynchronous versions of perfect recall, all admit decidable model checking in combination with quite rich temporal expressiveness [MS99, EGM07, HM10].

Orthogonally, a line of work on probabilistic model checking has considered model checking of assertions about probability and time [RKNP04]. Although one might at first expect this line of work to be closely related to epistemic model checking, in that probability theory provides a model of uncertainty, in fact this area has been concerned not with how subjective probabilities change over time, but with a probabilistic extension of temporal logic. The focus tends to be on the prior probability of some temporal property, or on the probability that some temporal property holds in runs from a current *known* state.

Rather less attention has been given to model checking the combination of subjective probability and temporal expressiveness. Of the semantics for knowledge mentioned above, the clock and synchronous perfect recall semantics are most suited as a basis for model checking subjective probability. (The others suffer from asynchrony, which makes it more difficult to associate a single natural probability space.) Implementations for these semantics presently exist only for a limited set of formulas, in which the full power of temporal logic is not used. For example, results in [HLM11] for model checking the logic of subjective probability (with clock or synchronous perfect recall semantics) and time restricts the temporal operators to have only finite reach into the future, and does not handle operators such as "at all times in the future".

A fundamental reason underlying this is that the problem of model checking probability with a rich temporal expressiveness seems to be inherently complex. Indeed, it requires a solution to a basic mathematical problem, the *Skolem Problem* for linear recurrences, that has stood unsolved since first posed in the 1930's [Sko34]. Consequently, the strongest results on model checking probability and time that encompass the expressiveness required for model checking knowledge and subjective probability state decidability in a way that requires exclusion of an infinite set of difficult instances for which decidability is unresolved. Specifically, [BRS06] shows that a (weak) monadic second order logic **PMLO**, containing probability assertions of forms such as $\texttt{Pr}(\phi(t_1, \ldots, t_n)) > c$, in which the t_i take values in the natural numbers, representing discrete time points, is decidable in finite state Markov chains, provided that the rational number c is not in a set H_ϕ depending on ϕ which can taken to be of arbitrarily small non-zero measure. This work leaves open the decidability of the model checking problem for the language in its full generality, in particular, for the values of c in H_ϕ.

Our contribution in this paper is to consider a number of generalizations of **PMLO**, motivated by model checking a logic of time and subjective probability. In particular, our generalizations arise very naturally when attempting to deal with the way that an agent conditions probability on its observations. We show that these generalizations definitively result in undecidable model checking problems. This clarifies the boundary between the decidable and undecidable cases of model checking logics of probability and time.

We begin in section 2 by recalling the definition of *probabilistic interpreted systems* [Hal03], which provides a very general semantic framework for logics of time, knowledge and probability. We work with an instantiation of this general framework in which systems are generated from finite state partially observed discrete-time Markov chains. We define two logics that take semantics in this framework. The first is an extension of the branching time temporal logic $\mathbf{CTL}^*$ to include operators for knowledge and probability, including operators for the subjective probability of agents. The second is a more expressive monadic second order logic that also adds a capability to quantify over moments of time and *finite sets* of moments of time. In this logic, the agent knowledge and probability operators are indexed by a temporal variable. This logic generalizes the logic of [BRS06]. Our logics allow polynomial comparisons of probability terms, as well as comparisons of agent probability terms referring to multiple time points. We argue from a number of motivating applications that this level of expressiveness is useful in potential applications. We show in Section 3 that the monadic second order logic is as least as expressive as our probabilistic extension of $\mathbf{CTL}^*$. Indeed, some apparently mild extensions of **PMLO** suffice for the encoding: the epistemic and subjective probability operators can be eliminated using a universal modality, polynomial combinations of probability expressions, and a more liberal use of quantification than allowed in **PMLO**.

We then turn in Section 4 to an investigation of the model checking problem. Specifically, we show that model checking even very simple formulas about a single agent's probability is undecidable when the agent has perfect recall. A consequence of this result is that an extension of **PMLO** that adds second order quantification into the scope of probability is undecidable.

This suggests a focus on weaker epistemic semantics instead, in particular, the clock semantics. From the point of view of **PMLO**, to express agent's subjective probabilities with respect to the clock semantics requires polynomial combinations of simple global probability terms of the form " the probability that proposition p holds at time t". We formulate a simple class of formulas involving such polynomial combinations, and show that this also has undecidable model checking.

These results show that even simple model checking questions about subjective probability are undecidable, and moreover help to explain some unexplained restrictions on **PMLO** in [BRS06]: these restrictions are in fact necessary in order to obtain a decidable logic. We conclude with a discussion of future work in Section 5. Related work most closely related to our results is discussed in the context of presenting and motivating the results.

2. PROBABILISTIC KNOWLEDGE

We describe in this section the semantic setting for the model checking problem we consider. We model a set of agents making partial observations of an environment that evolves with time. We first present the semantics of the modal logic we consider, following [Hal03], using the general notion of probabilistic interpreted system. Since these structures are not finite, in order to have a finite input for a model checking problem, we derive a probabilistic interpreted system from a partially observed discrete-time Markov chain. This is done in two ways, depending on the degree of recall of the agents. Taking the Markov chain to be finite, we obtain finitely presented model checking problems whose complexity we then study.

2.1 Probabilistic Interpreted Systems

Probabilistic interpreted systems are defined as follows. Let $Agt = \{1, \ldots, n\}$ be a set of agents operating in an environment e. At each moment of time, each agent is assumed to be in some *local* state, which records all the information that the agent can access at that time. The environment e records "everything else that is relevant". Let S be the set of environment states and let L_i be the set of local states of agent $i \in Agt$. A *global* state of a multi-agent system is an $(n+1)$-tuple $s = (s_e, s_1, \ldots, s_n)$ such that $s_e \in S$ and $s_i \in L_i$ for all $i \in Agt$. We write $\mathcal{G} = S \times L_1 \times \ldots \times L_n$ for the set of global states.

Time is represented discretely using the natural numbers $\mathbb{N}$. A *run* is a function $r : \mathbb{N} \to \mathcal{G}$, specifying a global state at each moment of time. A pair (r, m) consisting of a run r and time $m \in \mathbb{N}$ is called a *point*. If $r(m) = (s_e, s_1, \ldots, s_n)$ then we define $r_e(m) = s_e$ and $r_i(m) = s_i$ for $i \in Agt$. If r is a run and $m \in \mathbb{N}$ a time, we write $r[0..m]$ for $r(0) \ldots r(m)$ and $r_e[0..m]$ for $r_e(0) \ldots r_e(m)$. A *system* is a set $\mathcal{R}$ of runs. We call $\mathcal{R} \times \mathbb{N}$ the *set of points* of the system $\mathcal{R}$.

Agent knowledge is captured using a relation of indistinguishability. Two points (r, m) and (r', m') are said to be *indistinguishable to agent* i, if the agent is in the same local state at these points. Formally, we define $\sim_i$ to be the equivalence relation on $\mathcal{R} \times \mathbb{N}$ given by $(r, m) \sim_i (r', m')$, if $r_i(m) = r'_i(m')$. Relative to a system $\mathcal{R}$, we define the set

$$\mathcal{K}_i(r, m) = \{(r', m') \in \mathcal{R} \times \mathbb{N} \mid (r', m') \sim_i (r, m)\}$$

to be the set of points that are, for agent i, indistinguishable from the point (r, m). Intuitively, $\mathcal{K}_i(r, m)$ is the set of all points that the agent considers possible when it is in the actual situation (r, m). A system is said to be *synchronous* if for all agents i, we have that $(r', m') \in \mathcal{K}_i(r, m)$ implies that $m = m'$. Intuitively, in a synchronous system, agents always know the time. Since it is more difficult to define probabilistic knowledge in systems that are not synchronous, we confine our attention to synchronous systems in what follows.

A *probability space* is a triple $\mathbf{Pr} = (W, \mathcal{F}, \mu)$ such that W is a (nonempty) set, called the *carrier*, $\mathcal{F} \subseteq \mathcal{P}(W)$ is a σ-field of subsets of W, called the *measurable* sets in $\mathbf{Pr}$, closed under complementation and countable union, and $\mu : \mathcal{F} \to [0,1]$ is a *probability measure*, such that $\mu(W) = 1$ and $\mu(\bigcup_n V_n) = \sum_n \mu(V_n)$ for every countable sequence $\{V_n\}$ of mutually disjoint measurable sets $V_n \in \mathcal{F}$. As usual, we define the conditional probability $\mu(U|V) = \mu(U \cap V)/\mu(V)$ when $\mu(V) > 0$.

Let *Prop* be a set of *atomic propositions*. A *probabilistic interpreted system* over *Prop* is a tuple $\mathcal{I} = (\mathcal{R}, \mathtt{Pr}_1, \ldots, \mathtt{Pr}_n, \pi)$ such that $\mathcal{R}$ is a system, each $\mathtt{Pr}_i$ is a function mapping each point (r, m) of $\mathcal{R}$ to a probability space $\mathtt{Pr}_i(r, m)$ in which the carrier is a subset of $\mathcal{R} \times \mathbb{N}$, and $\pi : \mathcal{R} \times \mathbb{N} \to \mathcal{P}(Prop)$ is an interpretation of some set *Prop* of atomic propositions. Intuitively, the probability space $\mathtt{Pr}_i(r, m)$ captures the way that the agent i assigns probabilities at the point (r, m), and $\pi(r, m)$ is the set of atomic propositions that are true at the point.

We will work with probabilistic interpreted systems derived from synchronous systems in which agents have a common prior on the set of runs. To define these, we use the following notation. For a system $\mathcal{R}$, a set of runs $\mathcal{S} \subseteq \mathcal{R}$ and a set of points $U \subseteq \mathcal{R} \times \mathbb{N}$, define

$$\mathcal{S}(U) = \{r \in \mathcal{S} \mid \exists m : (r, m) \in U\}$$

to be the set of runs in $\mathcal{S}$ containing some point in the set U. Conversely, for a set $\mathcal{S}$ of runs and a set U of points, define

$$U(\mathcal{S}) = \{(r, m) \in U \mid r \in \mathcal{S}\}$$

to be the set of points in U that are on a run in $\mathcal{S}$. Note that if there exists a constant $k \in \mathbb{N}$ such that $(r, m) \in U$ implies $m = k$, then the relation $r \leftrightarrow (r, k)$ defines a one-to-one correspondence between $\mathcal{S}(U)$ and $U(\mathcal{S})$. In synchronous systems, which satisfy this condition, this gives a way to move between sets of points considered possible by an agent and corresponding sets of runs.

Suppose that $\mathcal{R}$ is a synchronous system, let $\mathbf{Pr} = (\mathcal{R}, \mathcal{F}, \mu)$ be a probability space on the system $\mathcal{R}$, and let π be an interpretation on $\mathcal{R}$. Intuitively, the probability space $\mathbf{Pr}$ represents a prior distribution over the runs. We assume that for all points $(r, m) \in \mathcal{R} \times \mathbb{N}$ and agents i, we have that $\mathcal{R}(\mathcal{K}_i(r, m)) \in \mathcal{F}$ is a measurable set and $\mu(\mathcal{R}(\mathcal{K}_i(r, m))) > 0$. (This assumption can be understood as saying that, according to the prior, each possible local state $r_i(m)$ of agent i at time m has non-zero probability of being the local state of agent i at time m.) Under this condition, we define the probabilistic interpreted system $\mathcal{I}(\mathcal{R}, \mathbf{Pr}, \pi) = (\mathcal{R}, \mathtt{Pr}_1, \ldots, \mathtt{Pr}_n, \pi)$ such that $\mathtt{Pr}_i$ associates with each point (r, m) the probability space $\mathtt{Pr}_i(r, m) = (\mathcal{K}_i(r, m), \mathcal{F}_{r,m,i}, \mu_{r,m,i})$ defined by

$$\mathcal{F}_{r,m,i} = \{\mathcal{K}_i(r, m)(\mathcal{S}) \mid \mathcal{S} \in \mathcal{F}\}$$

and such that

$$\mu_{r,m,i}(U) = \mu(\,\mathcal{R}(U) \mid \mathcal{R}(\mathcal{K}_i(r, m))\,)$$

for all $U \in \mathcal{F}_{r,m,i}$. Intuitively, because the set of runs $\mathcal{R}(\mathcal{K}_i(r, m))$ is measurable, we can obtain a probability space with carrier $\mathcal{R}(\mathcal{K}_i(r, m))$ by conditioning in $\mathbf{Pr}$. Because of the synchrony assumption there is, for each point (r, m), a one-to-one correspondence between points in $\mathcal{K}_i(r, m)$ and runs in $\mathcal{R}(\mathcal{K}_i(r, m))$. The construction uses this correspondence to induce a probability space on $\mathcal{K}_i(r, m)$ from the probability space on $\mathcal{R}(\mathcal{K}_i(r, m))$. We remark that under the additional assumption of perfect recall, it is also possible to understand each space $\mathtt{Pr}_i(r, m + 1)$ as obtained by conditioning on the space $\mathtt{Pr}_i(r, m)$. See [Hal03] for a detailed explanation of this point.

2.2 Probabilistic Temporal Epistemic Logic

To specify properties of probabilistic interpreted systems, a variety of logics can be formulated, drawing from the spectrum of temporal logics. Our main interest is in a reasoning about subjective probability and time, so we first consider a natural way to combine existing temporal and probabilistic logics. For purposes of comparison, it is also helpful to consider a rather richer monadic second order logic of probability and time, that is closely related to a logic for which some decidability results are known.

We may combine temporal and probabilistic logics to define a logic $\mathbf{CTL^*KP}$ that extends the temporal logic $\mathbf{CTL}^*$ by adding operators for knowledge and probability. Its syntax is given by the grammar

$$\phi \ ::= \ p \mid \neg\phi \mid \phi \wedge \phi \mid A\phi \mid X\phi \mid \phi U \phi \mid K_i\phi \mid f(P, \ldots, P) \bowtie c$$

$$P ::= \mathtt{Pr}_i(\phi) \mid \mathtt{Prior}_i(\phi)$$

where $p \in Prop$, c is a rational constant, $\bowtie$ is a relation symbol in the set $\{\leq, <, =, >, \geq\}$, and $f(x_1, \ldots, x_k)$ is multivariate polynomial in k variables $x_1, \ldots x_k$ with rational coefficients. Instances of P are called *basic probability expression.* The instances generated from $f(P, \ldots, P)$ are called *probability expressions*, and are expressions of the form $f(P_1, \ldots, P_k)$, obtained by substituting a basic probability expression P_i for each variable x_i in $f(x_1, \ldots, x_k)$. For example,

$$4\mathtt{Pr}_1(p)^5 \cdot \mathtt{Pr}_2(q)^3 + \frac{7}{15}\mathtt{Pr}_1(p)$$

is an instance of $f(P, \ldots, P)$ obtained from $f(x, y) = 4x^5y^3 + \frac{7}{15}x$ by substituting $\mathtt{Pr}_1(p)$ for x and $\mathtt{Pr}_2(q)$ for y.

Intuitively, formula $K_i\phi$ expresses that agent i knows ϕ. The formula $A\phi$ says that ϕ holds for all possible system evolutions from the current situation. The formula $X\phi$ expresses that ϕ holds at the next moment of time. The formula $\phi_1 U \phi_2$ says that ϕ_2 eventually holds, and ϕ_1 holds until that time. The expression $\mathtt{Pr}_i(\phi)$ represents agent i's current probability of ϕ, $\mathtt{Prior}_i(\phi)$ represents agent i's *prior* probability of ϕ, i.e., the agent's probability of ϕ at time 0. The formula $f(P_1, \ldots, P_k) \bowtie c$ expresses that this polynomial combination of current and prior probabilities stands in the relation $\bowtie$ to c. We use standard abbreviations from temporal logic, in particular, we write $F\phi$ for $true U\phi$.

A restricted fragment of the language that may be of interest is the *branching time fragment* in which the temporal operators are restricted to those of the temporal logic $\mathbf{CTL}$. That is, X and U are permitted to occur only in combination with the operator A, in one of the forms $AX\phi$, $EX\phi$, $A\phi_1 U \phi_2$, $E\phi_1 U \phi_2$, where we write $E\phi$ as an abbreviation for $\neg A \neg\phi$. We call this fragment of the language $\mathbf{CTLPK}$. The motivation for considering this fragment is that the complexity of model checking is in polynomial time for the temporal logic **CTL**, whereas it is polynomial-space complete for the richer temporal logic $\mathbf{CTL}^*$ [CES86]. The logic **CTLPK** is therefore, *prima facie*, a candidate for lower complexity once knowledge and probability operators are added to the logic.

The semantics of the language $\mathbf{CTL}^*\mathbf{KP}$ in a probabilistic interpreted system $\mathcal{I} = \mathcal{I}(\mathcal{R}, \mathbf{Pr}, \pi)$ is given by interpreting formulas ϕ at points (r,m) of $\mathcal{I}$, using a satisfaction relation $\mathcal{I},(r,m) \models \phi$. The definition is mutually recursive with a function $[\cdot]_{\mathcal{I},(r,m)}$ that assigns a value $[P]_{\mathcal{I},(r,m)}$ to each probability expression P at each point (r,m). This requires computing the measure of certain sets. For the moment, we assume that all sets arising in the definition are measurable. We show later that this assumption holds in the cases of interest in this paper.

We first interpret the probability expressions at points (r,m) of the system $\mathcal{I}$, by

$$[\mathtt{Pr}_i\phi]_{\mathcal{I},(r,m)} = \mu_{r,m,i}(\{(r',m') \in \mathcal{K}_i(r,m) \mid \mathcal{I},(r',m') \models \phi\})$$

$$[\mathtt{Prior}_i\phi]_{\mathcal{I},(r,m)} = \mu_{r,0,i}(\{(r',0) \in \mathcal{K}_i(r,0) \mid \mathcal{I},(r',0) \models \phi\})$$

$$[f(P_1,\ldots,P_k)]_{\mathcal{I},(r,m)} = f([P_1]_{\mathcal{I},(r,m)},\ldots,[P_k]_{\mathcal{I},(r,m)})$$

The satisfaction relation is then defined recursively, as follows:

1. $\mathcal{I},(r,m) \models p$ if $p \in \pi(r,m)$
2. $\mathcal{I},(r,m) \models \neg\phi$ iff not $\mathcal{I},(r,m) \models \phi$
3. $\mathcal{I},(r,m) \models \phi_1 \wedge \phi_2$ iff $\mathcal{I},(r,m) \models \phi_1$ and $\mathcal{I},(r,m) \models \phi_2$
4. $\mathcal{I},(r,m) \models A\phi$ if $\mathcal{I},(r',m) \models \phi$ for all runs r' with $r'[0\ldots m] = r[0\ldots m]$,
5. $\mathcal{I},(r,m) \models X\phi$ if $\mathcal{I},(r,m+1) \models \phi$
6. $\mathcal{I},(r,m) \models \phi_1 U \phi_2$ holds if there exists $k \geq m$ such that $\mathcal{I},(r,k) \models \phi_2$, and $\mathcal{I},(r,l) \models \phi_1$ for all l with $m \leq l < k$.
7. $\mathcal{I},(r,m) \models K_i\phi$ if $\mathcal{I},(r',m') \models \phi$ for all $(r',m') \in \mathcal{K}_i(r,m)$.
8. $\mathcal{I},(r,m) \models f(P_1,...,P_k) \bowtie c$ if $[f(P_1,...,P_k)]_{\mathcal{I},(r,m)} \bowtie c$.

2.3 Probabilistic Monadic Second Order Logic

Temporal modal logics refer to time in a somewhat implicit way. An alternative approach is to work in a setting with more explicit references to time, by using variables denoting time points. Kamp's theorem [Kam68] establishes an equivalence in the first order case, but by adding second order variables and quantification, one can obtain richer logics, that frequently remain decidable in the monadic case. In this section, we develop a logic in this style for time and subjective probability.

We define the logic **WMLOKP** as follows. We use two types of variables: time variables t and set variables X. Time variables take values in $\mathbb{N}$ and set variables take *finite* subsets of $\mathbb{N}$ as values. *Probability terms* P have the form $\mathtt{Pr}(\phi)$ or the form $\mathtt{Pr}_{i,t}(\phi)$ where $i \in Agt$ is an agent, t is a time variable, ϕ is a formula. Formulas ϕ are defined by the following grammar:

$$\phi ::= p(t) \mid X(t) \mid t_1 < t_2 \mid f(P,\ldots,P) \bowtie c \mid \neg\phi \mid \phi \wedge \phi \mid K_{i,t}(\phi) \mid \forall t(\phi) \mid \forall X(\phi)$$

where t, t_1, t_2 are time variables, p is an atomic proposition, X is a set variable, i is an agent, $c \in \mathbb{Q}$ is a rational constant, f is a rational polynomial (see the discussion above for $\mathbf{CTL}^*\mathbf{KP}$), and $\bowtie$ is a relation symbol from the set $\{=,<,\leq,>,\geq\}$.

Intuitively, in this logic formulas are interpreted relative to a run. Instead of indexing by a single moment of time, as in the logic above, we relativize the satisfaction relation to an assignment of values to the temporal and set variables. Atomic formula $p(t)$ says that proposition p holds at time t. Similarly, a (finite) set X of times can be interpreted as a proposition, and we can understand $X(t)$ as stating that the value of t is in X. (We remark that there is a fundamental difference between the types of propositions denoted by atomic propositions p and set variables X: whereas the atomic propositions may depend on structural aspects of the run, such as the global state at time t, the set variables may refer only to the time.) The atomic formula $t_1 < t_2$ has the obvious interpretation that time t_1 is less than time t_2. The constructs $\forall t(\phi)$ and $\forall X(\phi)$ correspond to universal quantification over times and *finite* sets of times respectively. They say that ϕ holds on the current run for all values of the variable. (Taking finite sets amounts to the *weak* interpretation of second order quantification. One could also consider a strong semantics allowing infinite sets of times. We have opted here for the weak interpretation to more easily relate our results to the existing literature.)

The probability term $\mathtt{Pr}(\phi)$ refers to the probability of ϕ in the probability space on runs. The meaning of probability term $\mathtt{Pr}_{i,t}(\phi)$ is agent i's probability at time t that the run satisfies ϕ. Similarly, $K_{i,t}\phi$ says that agent i knows at time t that the run satisfies ϕ. Note that, whereas in $\mathbf{CTL}^*\mathbf{KP}$, the formula $K_i\phi$ always expresses that agent i knows that ϕ holds at the "current time", in **WMLOKP**, formulas such as

$$\exists u(u < t \wedge K_{i,t}(p(u)))$$

talk about the agent's knowledge, at some time t, about what was true at some earlier time u. A similar point applies to probability expressions.

When dealing with formulas with free time and set variables, we need the extra notion of an assignment for the time and set variables. This is a function τ such that for each free time variable t we have $\tau(t) \in \mathbb{N}$, and for each free set variable X we have that $\tau(X)$ is a finite subset of $\mathbb{N}$. Given such an assignment, we give the semantics of probability terms and formulas by a mutual recursion. We give the semantics of formulas ϕ by means of a relation $\mathcal{I},\tau,r \models \phi$ defined as follows:

1. $\mathcal{I},\tau,r \models p(t)$ if $p \in \pi(r,\tau(t))$, when p is an atomic proposition
2. $\mathcal{I},\tau,r \models X(t)$ iff $\tau(t) \in \tau(X)$, if X is a set variable
3. $\mathcal{I},\tau,r \models \neg\phi$ iff not $\mathcal{I},\tau,r \models \phi$,
4. $\mathcal{I},\tau,r \models \phi_1 \wedge \phi_2$ iff $\mathcal{I},\tau,r \models \phi_1$ and $\mathcal{I},\tau,r \models \phi_2$
5. $\mathcal{I},\tau,r \models K_{i,t}(\phi)$ if $\mathcal{I},\tau,r' \models \phi$ for all $(r',m') \in \mathcal{K}_i(r,\tau(t))$.
6. $\mathcal{I},\tau,r \models f(P_1,...,P_k) \bowtie c$ if $[f(P_1,...,P_k)]_{\mathcal{I},\tau,r} \bowtie c$,
7. $\mathcal{I},\tau,r \models \forall t(\phi)$ if $\mathcal{I},\tau[t \mapsto n],r \models \phi$ for all $n \in \mathbb{N}$,
8. $\mathcal{I},\tau,r \models \forall X(\phi)$ if $\mathcal{I},\tau[X \mapsto U],r \models \phi$ for all finite $U \subseteq \mathbb{N}$.

In item (6), the definition is mutually recursive with the semantics of probability terms, which are interpreted as real numbers, relative to a temporal assignment. We define

$$[\mathtt{Pr}(\phi)]_{\mathcal{I},\tau,r} = \mu(\{r' \mid \mathcal{I},\tau,r' \models \phi\})$$

and

$$[\mathtt{Pr}_{i,t}(\phi)]_{\mathcal{I},\tau,r} = \frac{\mu(\{r' \mid (r,\tau(t)) \sim_i (r',\tau(t)),\ \mathcal{I},\tau,r' \models \phi\})}{\mu(\{r' \mid (r,\tau(t)) \sim_i (r',\tau(t))\})}$$

$$[f(P_1,\ldots,P_k)]_{\mathcal{I},\tau,r} = f([P_1]_{\mathcal{I},\tau,r},\ldots,[P_k]_{\mathcal{I},\tau,r})$$

As above, we assume measurability of the sets required, and later justify that this holds in the particular setting of interest in this paper.

A particular class of formulas of **WMLOKP** will be of interest below. Define a *mixed-time polynomial atomic probability formula* to be a formula of the form

$$\forall t_1 \ldots t_n (f(\mathtt{Pr}(\phi_1),\ldots,\mathtt{Pr}(\phi_m)) = 0)$$

where $f(x_1,\ldots,x_m)$ is a rational polynomial and each ϕ_i is an atomic formula of the form $p(t_j)$ for some proposition p and $j \in \{1\ldots n\}$. We motivate the usefulness of such temporal mixing of probability expressions in Section 2.5.

The logic **WMLOKP** generalizes several logics from the literature. If we restrict the language by excluding the probability comparison atoms $f(P_1,\ldots,P_k) \bowtie c$ and knowledge formulas $K_{i,t}(\phi)$, we have the *Weak Monadic Logic of Order*, which is equivalent to WS1S [Buc60]. We obtain the *Probabilistic Monadic Logic of Order* considered in [BRS06], which we denote here by **PMLO**, if we

- exclude the knowledge operators $K_{i,t}$,
- exclude agent's probability terms $\mathtt{Pr}_{i,t}(\phi)$, and
- limit the global probability comparisons to be of the form $\mathtt{Pr}(\phi(t_1,\ldots,t_k)) \bowtie c$, containing just a single probability term $\mathtt{Pr}(\phi(t_1,\ldots,t_k))$, with the further constraint that the only free variables of ϕ should be temporal variables $t_1,\ldots t_k$.

In particular, second-order quantification into probability expressions, e.g., $\forall X[\mathtt{Pr}(X(t)) > c]$ is not permitted in **PMLO**, but second order quantification that does not cross a probability operator, such as $\mathtt{Pr}(\forall X[X(t))]) > c$, is allowed. We note that **PMLO** *does* allow first order quantifications into the scope of probability, such as $\forall t[\mathtt{Pr}(p(t)) > c]$.

In the sequel, we refer to quantification into the scope of a knowledge formula or probability expression as *quantifying-in*.

2.4 Partially Observed Markov Chains

Although they provide a coherent semantic framework, probabilistic interpreted systems are infinite structures, and therefore not suitable as input for a model checking algorithm. We therefore work with a type of finite model called an *interpreted partially observed discrete-time Markov chain*, or PO-DTMC for short. A finite PO-DTMC for n agents is a tuple $M = (S, PI, PT, O_1, ..., O_n, \pi)$, where S is a finite set of states, $PI : S \to [0..1]$ is a function such that $\sum_{s\in S} PI(s) = 1$, component $PT : S \times S \to [0,1]$ is a function such that $\sum_{s'\in S} PT(s,s') = 1$ for all $s \in S$, and for each agent $i \in Agt$, we have a function $O_i : S \to \mathcal{O}$ for some set $\mathcal{O}$. Finally, $\pi : S \to \mathcal{P}(Prop)$ is an interpretation of the atomic propositions *Prop* at the states.

Intuitively, $PI(s)$ is the probability that an execution of the system starts at state s, and $PT(s,t)$ is the probability that the state of the system at the next moment of time will be t, given that it is currently s. The value $O_i(s)$ is the observation that agent i makes when the system is in state s. (Below, in the context of interpreted systems, we treat the set of states S as the states of the environment rather than as the set of global states. Agents' local states will be derived from the observations.)

Note that the first three components (S, PI, PT) of a PO-DTMC form a standard discrete-time Markov chain. This gives rise to a probability space on runs in the usual way. A *path* in M is a finite or infinite sequence $\rho = s_0 s_1 \ldots$ such that $PI(s_0) \neq 0$ and $PT(s_k, s_{k+1}) > 0$ for all k with $0 \le k < |\rho| - 1$. We write $\mathtt{P}_\infty(M)$ for the set of all infinite paths of M. Any finite path $\rho = s_0 s_1 \ldots s_m$ defines a set

$$\mathtt{P}_\infty(M) \uparrow \rho = \{\omega \in \mathtt{P}_\infty(M) \mid \omega[0\ldots m] = \rho\} \qquad (2)$$

That is, $\mathtt{P}_\infty(M) \uparrow \rho$ consists of all infinite paths which have ρ as a prefix.

We now define a probability space $\mathbf{Pr}(M) = (\mathtt{P}_\infty(M), \mathcal{F}, \mu)$ over the set $\mathtt{P}_\infty(M)$ of all infinite paths of M. The σ-algebra $\mathcal{F}$ is defined to be the smallest σ-algebra over $\mathtt{P}_\infty(M)$ that contains as basic sets all the sets $\mathtt{P}_\infty(M) \uparrow \rho$ for $\rho = s_0 s_1 \ldots s_m$ a finite path of M. For these basic sets, the function μ is defined by

$$\mu(\mathtt{P}_\infty(M) \uparrow \rho) = PI(s_0) \cdot PT(s_0, s_1) \cdot \ldots \cdot PT(s_{m-1}, s_m)\ .$$

The fact that μ can be extended to a measure on $\mathcal{F}$ is a non-trivial result of Kolmogorov for more general stochastic processes [KSK76].

We may construct several different probabilistic interpreted systems from each PO-DTMC, depending on what agents remember of their observations. We consider two, one that assumes that agents have perfect recall of all their observations, denoted $\mathtt{spr}$, and the other, denoted $\mathtt{clk}$, which assumes that agents are aware of the current time and their current observation. Recall that runs in an interpreted system map time to global states, consisting of a state of the environment and a local state for each agent. We interpret the states of the PO-DTMC M as states of the environment. To obtain a run, we also need to specify a local state for each agent at each moment of time. We use the the observations to construct the local states.

In the case of the *synchronous perfect recall semantics*, given a path $\rho \in \mathtt{P}_\infty(M)$, we obtain a run ρ^{spr} by defining the components at each time m as follows. The environment state at time m is $\rho_e^{\mathsf{spr}}(m) = \rho(m)$, and the local state of agent i at time m is $\rho_i^{\mathsf{spr}}(m) = O_i(\rho(0)) \ldots O_i(\rho(m))$. Intuitively, this local state assignment represents that the agent remembers all its past observations. We write $\mathcal{R}^{\mathsf{spr}}(M)$ for the set of runs of the form ρ^{spr} for $\rho \in \mathtt{P}_\infty(M)$. Note that this system is synchronous: if $r = \rho^{\mathsf{spr}}$ and $r' = \omega^{\mathsf{spr}}$ then for each agent i and time $m \in \mathbb{N}$, if $r_i(m) = r'_i(m')$, then $O_i(\rho(0)) \ldots O_i(\rho(m)) = O_i(\omega(0)) \ldots O_i(\omega(m'))$, which implies $m = m'$.

For the *clock semantics*, we construct a run a $\rho^{\mathtt{clk}}$ in which again the environment state at time m is $\rho_e^{\mathtt{clk}}(m) = \rho(m)$, and for agent i we define the local state at time m by $\rho^{\mathtt{clk}}(m) = (m, O_i(\rho(m)))$. Intuitively, this says that the agent is aware of the clock value and its current observation. We write $\mathcal{R}^{\mathtt{clk}}(M)$ for the set of runs of the form $\rho^{\mathtt{clk}}$ for $\rho \in \mathtt{P}_\infty(M)$ an infinite path of M. This system is also synchronous: if $r = \rho^{\mathtt{clk}}$ and $r' = \omega^{\mathtt{clk}}$ then for each agent i and time $m \in \mathbb{N}$, if $r_i(m) = r'_i(m')$, then $(m, O_i(\rho(m))) = (m', O_i(\omega(m')))$, hence $m = m'$. In both cases of $x \in \{\mathtt{spr}, \mathtt{clk}\}$, if T is a subset of $\mathtt{P}_\infty(M)$, we write T^x for $\{\rho^x \mid \rho \in T\}$.

In both cases of $x \in \{\mathtt{spr}, \mathtt{clk}\}$, we have a one-to-one correspondence between the infinite paths $\mathtt{P}_\infty(M)$ and the runs $\mathcal{R}^x(M)$. We therefore can induce probability spaces $\mathbf{Pr}^x(M)$ on $\mathcal{R}^x(M)$ from the probability space $\mathbf{Pr}(M)$ on $\mathtt{P}_\infty(M)$. As described above, the probability space $\mathbf{Pr}^x(M)$ on runs moreover induces a probability space $\mathtt{Pr}^x_i(r, m)$ on the set of points considered possible by each agent i at each point (r, m). The PO-DTMC M gives us an interpretation π on its states, and we may derive from this an interpretation π^x on the points (r, m) of $\mathcal{R}^{\mathtt{spr}}(M)$ and $\mathcal{R}^{\mathtt{clk}}(M)$ by defining $\pi^x(r, m) = \pi(r_e(m))$. Using the general construction defined above, we then obtain the probabilistic interpreted systems $\mathcal{I}^x(M) = \mathcal{I}(\mathcal{R}^x(M), \mathbf{Pr}^x(M), \pi^x)$ for $x \in \{\mathtt{spr}, \mathtt{clk}\}$.

It is necessary to establish the measurability of the sets corresponding to formulas for the semantic definitions of the logics above to be complete. This is established in the following result.

Lemma 1. *Let M be a finite PO-DTMC and $x \in \{\mathtt{spr}, \mathtt{clk}\}$. For every set $S \subseteq \mathcal{R}(M)$ of runs of M such that the semantic definitions above of* **CTL*KP** *and* **WMLOKP** *in $\mathcal{I}^x(M)$ refer to $\mu(S)$, the set S is measurable in $\mathbf{Pr}(M)$.*

2.5 Discussion

We have defined our logics to be quite expressive in the type of atomic probability assertions we have allowed, which involve polynomials of probability expressions. In **WMLOKP**, these expressions may explicitly refer to different time points. Some existing logics of probability in the literature use a more restricted expressiveness, e.g., [FH94] consider a logic that has only linear combinations of probability expressions, and many logics [BRS06, RKNP04] allow only inequalities involving a single probability term. Here give some motivation to show that the richness we have allowed is natural and useful for applications.

Polynomials: There are several motivations for allowing polynomial combinations of probability expressions. One, as noted in [FHM90], is that polynomials arise naturally from conditional probability. If we would like to include linear combinations of conditional probability expressions in the language, we find that this motivates a generalization to polynomial combinations of probability expressions. Consider the formula $\mathtt{Pr}(\phi_1|\psi_1) + \mathtt{Pr}(\phi_2|\psi_2) \leq c$. Expanding out the definition of conditional probability, we have

$$\frac{\mathtt{Pr}(\phi_1 \wedge \psi_1)}{\mathtt{Pr}(\psi_1)} + \frac{\mathtt{Pr}(\phi_2 \wedge \psi_2)}{\mathtt{Pr}(\psi_2)} \leq c \; .$$

We see here that there is a risk of division by zero that needs to be managed in order for the semantics of this formula to be fully defined. One way to do so is to multiply out the denominators, resulting in the form

$$\mathtt{Pr}(\phi_1 \wedge \psi_1) \cdot \mathtt{Pr}(\psi_2) + \mathtt{Pr}(\phi_2 \wedge \psi_2) \cdot \mathtt{Pr}(\psi_1) \leq c \cdot \mathtt{Pr}(\psi_1) \cdot \mathtt{Pr}(\psi_2)$$

which is meaningful in all cases. (Should this not have the desired semantics in case one of the $\mathtt{Pr}(\psi_i)$ is zero, an additional formula can be added that handles this special case as desired.) However, although we started with a linear probability expression, we now have multiplicative terms. This suggests that the appropriate way to add the expressiveness of conditional probability to the language is to admit atomic formulas that compare polynomial combinations of probability expressions.

More generally, although it is less of relevance for purposes of model checking, and more of use for axiomatization of the logic, allowing polynomials also naturally enables familiar reasoning patterns to be captured inside the logic. In particular, validities such as $\mathtt{Pr}(\phi_1 \vee \phi_2) = \mathtt{Pr}(\phi_1) + \mathtt{Pr}(\phi_2)$ when ϕ_1 and ϕ_2 are mutually exclusive and $\mathtt{Pr}(\phi_1 \wedge \phi_2) = \mathtt{Pr}(\phi_1) \cdot \mathtt{Pr}(\phi_2)$ when ϕ_1 and ϕ_2 are independent show that both addition and multiplication of probability terms arises naturally.

Mixed-time: A second way in which our logics are rich is in allowing probability atoms that refer to different moments of time. In **CTL*KP** this already the case because combinations such as $\mathtt{Prior}_A(\phi) = \mathtt{Pr}_A(\phi)$ are allowed, which refer to both the current time and to time 0. The logic **WMLOKP** takes such temporal mixing further by allowing reference to time points explicitly named using time variables.

Such temporal mixing is natural, since there are potential applications that require this expressiveness. For example, in computer security, one often wants to say that the adversary A does not learn anything about a secret from watching an exchange between two parties. However, it is often the case that the adversary knows some prior distribution over the secrets. (For example, the secret may be a password, and choice of passwords by users are very non-uniform, with some passwords like '123456' having a very high probability.) This means that the simple assertion that the adversary does not know the secret, or that the adversary has a uniform distribution over the secret, does not capture the appropriate notion of security. Instead, as recognised already by Shannon in his work on secrecy [Sha49], we need to assert that the adversary's distribution over the secret has not changed as a result of its observations. This requires talking about the adversary's probability at two time points. For example, [HLM11] capture an anonymity property by means of formulas using terms $\mathtt{Prior}_A(\phi) = \mathtt{Pr}_A(\phi)$.

Mixed-time polynomials: Additionally, the logic of probability applied to formulas referring to different times leads naturally to polynomial combinations of probability terms, each referring to a different moment of time. For example, although **PMLO** allows only formulas of the form $\mathtt{Pr}(\phi(t_1, \ldots, t_n)) \bowtie c$, where the t_i are time variables, the decision algorithm of [BRS06] uses the fact that, when $t_1 < t_2 < \ldots < t_n$, the formula $\phi(t_1, \ldots, t_n)$ is equivalent to a formula of the form $\phi_1(t_1) \wedge \phi_2(t_2 - t_1) \wedge \ldots \phi_n(t_n - t_{n-1}) \wedge \phi_{n+1}(t_n)$, where the $\phi_i(t)$ are independent past-time formulas for $i = 1 \ldots n$ and $\phi_{n+1}(t)$ is a future time formula. (This statement is closely related to Kamp's theorem [Kam68].) This enables $\mathtt{Pr}(\phi(t_1, \ldots, t_n))$ to be expressed as a sum of products of terms of the form $\mathtt{Pr}(\phi_i(u))$ where $\phi_i(u)$ has just a single free time variable u. Thus, although mixed-time probability formulas are not directly expressible in the logic of [BRS06], specific ones are implicitly expressible, and the extension is a mild one. It is worth remarking, however, that the coefficients of the polynomial expansion of $\mathtt{Pr}(\phi(t_1, \ldots, t_n))$ are all positive, so we do not quite have arbitrary polynomials here. We return to this point below.

3. RELATING THE LOGICS

The logic **WMLOKP** is very expressive, so it is not surprising that it can capture all of **CTL*KP**. The following result makes this precise.

For the results below, it is convenient to add to the system a special agent $\bot$ that is blind, and an agent $\top$ that that has complete information about the state. In the context of PO-DTMC's these agents are obtained by taking the observation functions to satisfy $O_\bot(s) = O_\bot(t)$ and $O_\top(s) = s$ for all states s, t. We write $\Box\phi$ for $K_{\bot,t}\phi$ where t is any time variable. This gives a *universal modality*: $\Box\phi$ says that ϕ holds on all runs.

Proposition 2. *Let M be a PO-DTMC with agent $\top$ and let $x \in \{\mathtt{spr}, \mathtt{clk}\}$. For every formula ϕ of* **CTL*KP***, there exists a formula $\phi^*(t)$ of* **WMLOKP** *with t the only free variable, such that $\mathcal{I}^x(M), (r, n) \models \phi$ iff $\mathcal{I}^x(M), [t \mapsto n], r \models \phi^*(t)$ for all runs r.*

Proof. The translation is defined by the following recursion:

$$p^*(t) = p(t),\ (\neg\phi)^*(t) = \neg\phi^*(t),\ (\phi_1 \wedge \phi_2)^*(t) = \phi_1^*(t) \wedge \phi_2^*(t),$$
$$(X\phi)^*(t) = \exists u(u = t+1 \wedge \phi^*(u)),\ (K_i\phi)^*(t) = K_{i,t}(\phi^*(t)),$$
$$(\phi_1 U \phi_2)^*(t) = \exists u \geq t(\phi_2^*(u) \wedge \forall v(t \leq v < u \Rightarrow \phi_1^*(v)))$$
$$(\mathtt{Pr}_i(\phi))^*(t) = \mathtt{Pr}_{i,t}(\phi^*(t)),\ (\mathtt{Prior}_i(\phi))^*(t) = \mathtt{Pr}_{i,0}(\phi^*(0))$$
$$(f(P_1, \ldots, P_k) \bowtie c)^*(t) = f(P_1^*(t), \ldots, P_k^*(t)) \bowtie c$$

Note that $u = t + 1$ is definable as $(u > t \wedge \forall v > t(u \leq v))$, and $u = 0$ is definable as $\neg\exists t(u = t + 1)$. We can use $(A\phi)^*(t) = K_{\top,t}(\phi^*(t))$ to translate $A\phi$ in the perfect recall case. In case of the clock semantics, this translation loses the information about the initial state, which is required for correctness of the translation of $\mathtt{Prior}_i(\phi)$. In this case, we introduce without loss of generality new propositions p_s for each state s, such that $p_s \in \pi_e(t)$ iff $s = t$, and take

$$(A\phi)^*(t) = \bigwedge_{s \in S}(p_s(0) \Rightarrow K_{\top,t}(p_s(0) \Rightarrow \phi^*(t)))$$

□

With respect to the specific systems we derive from PO-DTMC's with respect to the clock and perfect recall semantics, we are able to make some further statements that simplify the logic **WMLOKP** by eliminating some of the operators. These results are useful for the undecidability results that follow.

For the following results, we note that, without loss of generality, we may assume that a finite PO-DTMC comes equipped with atomic propositions that encode the observations made by the agents. Specifically, when agent i has possible observations $o_{i,1}, \ldots, o_{i,k_i}$, we assume that there are atomic propositions $obs_{i,j}$ for $i \in Agt$ and $j = 1 \ldots k_i$ such that for all states s, $obs_{i,j} \in \pi(s)$ iff $O_i(s) = o_{i,j}$. Thus, $obs_{i,j}(t)$ holds in a run just when agent i makes observation $o_{i,j}$ at time t.

Proposition 3. *With respect to $\mathcal{I}^{\mathtt{clk}}(M)$ for a finite PO-DTMC M, the operators $K_{i,t}$ and $\mathtt{Pr}_{i,t}$ can be eliminated using the universal operator $\Box$ and polynomial comparisons of universal probability terms $\mathtt{Pr}(\psi)$, respectively. For simple probability formulas $\mathtt{Pr}_{i,t}(\phi) \bowtie c$, only linear probability comparisons are required.*

Proof. The formula

$$\bigwedge_{j=i\ldots k_i}(obs_{i,j}(t) \Rightarrow \Box(obs_{i,j}(t) \Rightarrow \phi)$$

is easily seen to be equivalent to $K_{i,t}(\phi)$ in $\mathcal{I}^{\mathtt{clk}}(M)$. Similarly, $\mathtt{Pr}_{i,t}(\phi) \bowtie c$ can be expressed as

$$\bigwedge_{j=i\ldots k_i}(obs_{i,j}(t) \Rightarrow \mathtt{Pr}(obs_{i,j}(t) \wedge \phi) \bowtie c \cdot \mathtt{Pr}(obs_{i,j}(t)))\ .$$

A similar transformation applies for more general agent probability comparisons, but we note that linear comparisons may transform to polynomial comparisons: similarly to the discussion of conditional probability in Section 2.5. □

Proposition 4. *With respect to $\mathcal{I}^{\mathtt{spr}}(M)$ for a finite PO-DTMC M, the probability formulas $\mathtt{Pr}_{i,t}(\phi) \bowtie c$ can be reduced to linear comparisons using only terms $\mathtt{Pr}(\psi)$, provided second-order quantifying-in is permitted. Knowledge terms $K_{i,t}$ can be reduced to the universal modality $\Box$, provided second-order quantifying-in is permitted for this modality.*

Proof. Define $\kappa_i(X_1, \ldots, X_{k_i}, t)$ to be the formula

$$\forall t' \leq t(\bigwedge_{j=1\ldots k_i} X_i(t') \Leftrightarrow obs_{i,j}(t'))$$

Intuitively, this says that, up to time t, the second order variables $X_1, \ldots, X_k$ encode the pattern of occurrence of observations of agent i up to time t. The formula

$$\forall X_1, \ldots X_{k_i}(\kappa_i(X_1, \ldots, X_{k_i}, t) \Rightarrow \Box(\kappa_i(X_1, \ldots, X_{k_i}, t) \Rightarrow \phi)$$

is easily seen to be equivalent to $K_{i,t}(\phi)$ in $\mathcal{I}^{\mathtt{clk}}(M)$. Similarly, $\mathtt{Pr}_{i,t}(\phi) \bowtie c$ can be expressed as

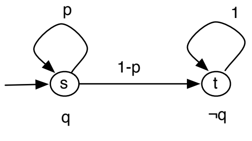
$$\forall X_1, \ldots X_{k_i}(\quad \kappa_i(X_1, \ldots, X_{k_i}, t) \Rightarrow$$
$$\mathtt{Pr}(\kappa_i(X_1, \ldots, X_{k_i}, t) \wedge \phi) \bowtie c \cdot \mathtt{Pr}(\kappa_i(X_1, \ldots, X_{k_i}, t)))$$

□

One might wonder whether the knowledge operators can be eliminated entirely using probability, treating $K_i\phi$ as $\mathtt{Pr}_i(\phi) = 1$. This is indeed the case for formulas ϕ in **CTLPK**. The essential reason is that because formulas of **CTLPK** depend at a point (r, m) only on the run prefix $r[0 \ldots m]$, so the possibility that $\neg\phi$ holds on a non-empty set of measure zero does not occur.

Proposition 5. *For all* **CTLPK** *formulas ϕ and PO-DTMC's M and $x \in \{\mathtt{clk}, \mathtt{spr}\}$ we have $\mathcal{I}^x(M) \models K_i\phi \Leftrightarrow \mathtt{Pr}_i(\phi) = 1$.*

However, this is not the case for formulas $K_i\phi$ where ϕ is an LTL formula. Consider the following Markov Chain. Here we have, at the initial state s, that $\neg K_i(F\neg q)$, because the agent considers it possible that always q (this holds for all choices of observation functions). However, we have $\mathtt{Pr}_i(F\neg q) = 1$, since the only run where $\neg q$ does not eventually hold is the run that always remains at s. This run has probability zero.



4. UNDECIDABILITY RESULTS

We can now state the main results of the paper concerning the problem of model checking formulas of (fragments of) the logics **CTL*KP** and **WMLOKP** in a PO-DTMC M, with respect to an epistemic semantics $x \in \{\texttt{spr}, \texttt{clk}\}$. Using the results of Section 3, we also obtain conclusions about extensions of **PMLO** that do not refer to agent probability and knowledge.

For a formula ϕ of **CTL*KP**, we write $M \models^x \phi$, if $\mathcal{I}^x(M), (r, 0) \models \phi$ for all runs $r \in \mathcal{R}^x(M)$. In the case of **WMLOKP**, we consider sentences, i.e., formulas without free variables, and write $M \models^x \phi$, if $\mathcal{I}^x(M), \tau, r \models \phi$ for all runs $r \in \mathcal{R}^x(M)$ and the empty assignment τ. The model checking problem is to determine, given a PO-DTMC M, a formula ϕ, and semantics $x \in \{\texttt{clk}, \texttt{spr}\}$, whether $M \models^x \phi$.

4.1 Background

For comparison with results below, it is worth stating a result from [BRS06] concerning decidability of the fragment **PMLO** of **WMLOKP** that omits knowledge operators $K_{i,t}$ and agent probability terms $\texttt{Pr}_{i,t}(\phi)$, restricts probability comparisons to the form $\texttt{Pr}(\phi) \bowtie c$, and prohibits second order quantification to cross into probability terms. Since the structure of agent's local states is irrelevant in this case, we write simply $\mathcal{I}(M)$ for the probabilistic interpreted system corresponding to a PO-DTMC M. To state the result, we define the *parameterized* variant of a formula ϕ of **PMLO** to be the formula $\phi_{x_1,\ldots,x_k}$, in which each subformula of the form $\texttt{Pr}(\psi) \bowtie c$ is replaced by a formula $\texttt{Pr}(\psi) \bowtie x_i$, with x_i a fresh variable. We call the resulting formulas the *parameterized formulas of* **PMLO**. For some $\alpha \in \mathbb{Q}^k$, we can then recover the original formula ϕ as the instance ϕ_α obtained from the parameterized variant $\phi_{x_1,\ldots,x_k}$ of ϕ by substituting α_i for x_i for each $i = 1 \ldots k$.

Theorem 6 ([BRS06]). *For each parameterized sentence $\phi_{x_1,\ldots,x_k}$ of* **PMLO***, one can compute for all $\epsilon > 0$ a representation of a set $H_\phi \subset \mathbb{R}^k$ of measure at most ϵ, such that the problem of determining if $\mathcal{I}(M) \models \phi_\alpha$ is decidable for $\alpha \in \mathbb{Q} \setminus H_\phi$.*

Intuitively, the complement of H_ϕ contains the points that are bounded away from limit points of the Markov chain, and comparisons can be decided using convergence properties.

The reason for excluding the set H_ϕ is that the limit point cases seem to require a resolution of problems related to the *Skolem problem* concerning zeros of linear recurrences [Sko34]. A sequence of real numbers $\{u_n\}$ is called a linear recurrence sequence (LRS) of order k if there exist $a_1, \ldots a_k$ with $a_k \neq 0$ such that for all $m \geq 1$,

$$u_{k+m} = a_1 u_{k+m-1} + a_2 u_{k+m-2} + \cdots + a_k u_m \ .$$

We consider the following decision problems associated with a LRS $\{u_n\}$.

1. **Skolem problem.** Does there exist n such that $u_n = 0$?

2. **Positivity problem.** Is it the case that for all n, $u_n \geq 0$?

3. **Ultimate positivity problem.** Does a positive integer N exist such that for all $n \geq N$, $u_n \geq 0$?

We will deal with sequences with rational entries. By clearing denominators the rational version of the above problems can be shown to be polynomially equivalent to similar problems stated using sequences with integer entries. There has been a significant amount of work on these problems [ESPW03], but they have stood unresolved since the 1930's. To date, only low order versions of these problems have been shown to be decidable [HHHK05, OW14, TMS84].

The above problems have an equivalent matrix formulation. A proof of the following can be found in [HHHK05].

Lemma 7. *For a sequence $u_0, u_1, \ldots,$ the following are equivalent.*

1. *$\{u_n\}$ is a rational LRS.*

2. *For $n \geq 1$, $u_n = (A^n)_{1k}$ for a square matrix A with rational entries.*

3. *For $n \geq 1$, $u_n = \mathbf{v}^T A^n \mathbf{w}$ where A is a square matrix, and $\mathbf{v}$ and $\mathbf{w}$ are vectors with entries from $\{0, 1\}$.*

In the usual formulation of the Skolem, positivity and ultimate positivity problems, the associated matrices A may contain negative numbers, and numbers not in $[0, 1]$, so are not stochastic matrices. However, [AAOW15] show that these problems can be reduced to a decision problem stated with respect to stochastic matrices:

Lemma 8. *Given an integer $k \times k$ matrix A, one can compute a $k' \times k'$ stochastic matrix B, a length k' stochastic vector $\mathbf{v}$, a length k' vector $\mathbf{w} = (0, \ldots, 0, 1)$ and a constant c such that $(A^n)_{1,k} = 0$ ($(A^n)_{1,k} > 0$) iff $\mathbf{v}^T B^n \mathbf{w} = c$ (respectively, $\mathbf{v}^T B^n \mathbf{w} > c$).*

As noted in [AAOW15], it follows that the logic **PMLO** is able to express the Skolem and positivity problems, using model checking questions of the form

$$M \models \exists t(\texttt{Pr}(p(t)) = c)$$

and

$$M \models \exists t(\texttt{Pr}(p(t)) > c)$$

for c a nonzero constant and M a DTMC. (The ultimate positivity problem can also be expressed.) It is worth noting that in the special case of the constant $c = 0$, these model checking questions *are* decidable, as shown in [BRS06]. (Essentially, in this case the problems reduce to graph reachability problems, and the specific probabilities in M are irrelevant.) The transformation from arbitrary matrices A to stochastic matrices B in Lemma 8 requires that the constant 0 of the Skolem problem be replaced by a non-zero constant c.

The above model checking problems of the quantified logic **PMLO** can be seen to be already expressible in the propositional logic **CTL*KP**, as the problems

$$M' \models^{\texttt{clk}} \mathbf{AF}(\mathrm{pr}_i(p) = c)$$

$$M' \models^{\texttt{clk}} \mathbf{AF}(\mathrm{pr}_i(p) > c)$$

$$M' \models^{\texttt{clk}} \mathbf{AFAG}(\mathrm{pr}_i(p) > c)$$

where we obtain the PO-DTMC M' from the DTMC M by defining $O_i(s) = \bot$ for all states s. That is, agent i is blind, so considers all states reachable at time n to be possible at time n. (We remark that this implies that all the operators A can be exchanged with E without change of meaning of the formulas.) It follows, that with respect to clock semantics, a resolution of the decidability of model checking even these simple formulas of **CTL*KP** for *all* $c \in [0,1]$ would imply a resolution of the Skolem problem. In view of the effort already invested in the Skolem problem, this is likely to be highly nontrivial.

4.2 Perfect Recall Semantics

Model checking with respect to the perfect recall semantics is undecidable, even with respect to a very simple fixed formula of the logic **CTL*KP**, as shown by the following result.

THEOREM 9. *The problem of determining, given a PO-DTMC M, if $M \models^{\mathrm{spr}} EF(\mathtt{Pr}_i(p) > c)$, for p an atomic proposition, is undecidable.*

PROOF. (Sketch) By reduction from the emptiness problem for probabilistic finite automata [Paz71]. Intuitively, the proof sets up an association between words in the matrix semigroup and sequences of observations of the agent.

A probabilistic finite automaton is a tuple $\mathcal{A} = (Q, \Sigma, \mathbf{v}_0, A, F, \lambda)$, where Q is a finite set of states, Σ is a finite alphabet, $\mathbf{v}_0 : Q \to [0,1]$ is a probability distribution over states, representing the initial distribution, $A : \Sigma \to (Q \times Q \to [0,1])$ associates a transition probability matrix $A(a)$ with each letter $a \in \Sigma$, component $F \subseteq Q$ is a set of *final* states, and $\lambda \in (0,1)$ is a rational number. Each matrix $A(a)$ satisfies $\sum_{t \in S} A(a)(s,t) = 1$ for all $s \in Q$. Let v_F be the column vector indexed by Q with $v_F(s) = 0$ if $s \notin F$ and $v_F(s) = 1$ if $s \in F$. Treating $\mathbf{v}_0$ as a row vector, for each word $w = a_1 \ldots a_n \in \Sigma^+$, define $f(w) = \mathbf{v}_0 A(a_1) \ldots A(a_n) v_F$. The language accepted by the automaton is defined to be $\mathcal{L}(\mathcal{A}) = \{w \in \Sigma^+ \mid f(w) > \lambda\}$. The emptiness problem for probabilistic finite automata is then, given a probabilistic finite automaton $\mathcal{A}$, to determine if the language $\mathcal{L}(\mathcal{A})$ is empty. This problem is known to be undecidable [Paz71, CL04].

Given a probabilistic finite automaton $\mathcal{A}$, we construct an interpreted finite PO-DTCM $M_{\mathcal{A}}$ for a single agent (called i rather than 1 to avoid confusion with other numbers) such that $M_{\mathcal{A}} \models^{\mathrm{spr}} EF(\mathtt{Pr}_i(p) > \lambda)$ iff $\mathcal{A}$ is nonempty. This system is defined as follows. We let $N = |\Sigma|$,

1. $S = Q \times \Sigma$,
2. $PI(q,a) = \mu_0(q)/N$,
3. $PT((q,a),(q',b)) = A(b)(q,q')/N$
4. $O_i((q,a)) = a$
5. $p \in \pi((q,a))$ iff $q \in F$.

Note that $\sum_{(q,a)\in S} PI((q,a)) = \sum_{a\in\Sigma}\sum_{q\in Q} \mu_0(q)/N = \sum_{a\in\Sigma} 1/N = 1$, so PI is in fact a distribution. Similarly, for each $(q,a) \in S$, we have $\sum_{(q',b)\in S} PT((q,a),(q',b)) = \sum_{b\in\Sigma}\sum_{q'\in Q} A(b)(q,q')/N = \sum_{b\in\Sigma} 1/N = 1$, so PT is in fact a stochastic matrix.

Note that for each $w = a_1 \ldots a_n \in \Sigma^*$ and $a \in \Sigma$, we get a row vector $\mu_{aw} = \mu_0 A(a_1) \ldots A(a_n)$ with $\sum_{q\in Q} \mu_{aw}(q) = 1$, which can be understood as a distribution on Q. For each run $r \in \mathcal{R}^{\mathrm{spr}}(M_{\mathcal{A}})$ and $m \geq 0$, we have that that agent i's local state $r_i[0 \ldots m]$ at (r,m) is a word in Σ^+. Let $\mathcal{B}(q,m)$ be the set of runs $r \in \mathcal{R}^{\mathrm{spr}}(M_{\mathcal{A}})$ in which $r_e(m) = (q,a)$ for some $a \in \Sigma$. We claim the following about the probability measures $\mu_{r,m,i}$ in the probabilistic interpreted system $\mathcal{I}^{\mathrm{spr}}(M_{\mathcal{A}})$, for each point (r,m) and $q \in Q$:

$$\mu_{r,m,i}(\mathcal{K}_i(r,m)(\mathcal{B}(q,m))) = (\mathbf{v}_0 A(r_i(1)) \ldots A(r_i(m)))(q) \ .$$

It is immediate from this that $\mathcal{I}^s pr(M_{\mathcal{A}}), (r,m) \models \mathtt{Pr}_i(p) = c$ where $c = f(r[1 \ldots m])$, and the desired result follows. □

We remark that this result stands in contrast to the situation for model checking the logic of knowledge and time. Write **CLTL*K** for the logic obtained from **CTL*KP** by omitting the probability comparison atoms $f(P_1, \ldots, P_k) \bowtie c$. Model checking the logic **CLTL*K** with respect to perfect recall, i.e., deciding $M \models^{\mathrm{spr}} \phi$ for M a PO-DTMC and ϕ a formula is decidable [MS99]. (Here, for the semantic structures M, it suffices to replace the initial distribution PI in M by the set $I = \{s \in S \mid PI(s) > 0\}$, and replace the transition distribution function PT in M by the relation R of possibility of transitions between states defined by sRt if $PT(s,t) > 0$. The results in [MS99] use linear time temporal logic as a basis, but, as noted in [MW03], the modality A of the branching time logic CTL^* can be understood as a special case of a knowledge modality: see Proposition 2.)

For probabilistic automata the minimum size of the state space giving undecidability directly stated in the literature appears to be 25 [Hir06]. We remark that the proof of Theorem 9 can also be done by reduction of the following matrix semigroup problem: *given a finite set of matrices of order n, generating a matrix semigroup S, determine whether there is $M \in S$ such that* $(M)_{1n} = 0$ [Hal97]. The case of k generators of size $n \times n$ can be reduced to probabilistic automata with $2kn + 1$ states. Recent results on the matrix semigroup problem are given in [CHHN14].

Huang et al [HSZ12] have previously used a reduction from probabilistic automata to show undecidability of an probabilistic epistemic logic with respect to perfect recall. Compared to our simple CTL temporal operators, their logic uses more expressive setting of alternating temporal logic operators.

4.3 Clock Semantics

The undecidability of the perfect recall semantics for such simple formulas suggests that we weaken the epistemic semantics to the clock case. The combination of the translation from **CTL*KP** to **WMLOKP** (Proposition 2) and Theorem 6 then enables some cases of **CTL*KP** to be decided. We do not obtain a full decidability result, however, since we face the problem that, with respect to the clock semantics, the formula $AF(\mathtt{Pr}_i(p) = c)$ can express the Skolem problem, so resolving its decidability is a very difficult problem. Rather than attempt to resolve this question, we consider here just how much extra expressiveness is required over the logic of Theorem 6 for us to obtain a definitive *undecidability* result, instead of a decidability result with some excluded and unresolved cases.

Note that one of the restrictions on **PMLO** used in Theorem 6 is that second order quantification should not cross into probability terms. It turns out that this restriction is essential, as shown by the following result.

THEOREM 10. *It is is undecidable, given a PO-DTMC M and a formula ϕ of* **WMLOKP** *with linear combinations of* *probability terms* $\mathtt{Pr}(\phi)$ *and quantifying-in of second-order quantifiers, whether* $M \models \phi$.

Proof. This follows from the fact that, using second order quan-tifying-in, we can express perfect recall (Proposition 4), and the undecidability of model checking perfect recall (Theorem 9). □

Note that the result refers to $\models$ rather than $\models^{\mathtt{clk}}$, since epistemic operators are not required. This is really a result about a generalization of **PMLO**. One of the other restrictions in Theorem 6 is that only simple probability comparisons of the form $\mathtt{Pr}(\phi) \bowtie c$ are permitted. More general comparisons of probability terms are needed in applications (see discussion in Section 2.5), so it is of interest to study their impact on decidability. Unfortunately, it turns out to be quite negative. Even the simple case of mixed time polynomial atomic probability formulas is enough for undecidability.

Theorem 11. *It is undecidable, given a DTMC* M *and a mixed-time polynomial atomic probability formula* ψ*, whether* $M \models \psi$.

Proof. (Sketch) By reduction from Hilbert's tenth problem, i.e., the problem of determining whether a polynomial with integer coefficients has solutions in the natural numbers. This was shown to be undecidable by Matiyasevich [Mat93].

Write e_j for a basic column vector (of the appropriate dimension) with 1 in the jth place and 0 elsewhere. Given a PO-DTMC with transition matrix M, with a proposition for each state, we can express the values $\mathbf{e}_i^T M^t \mathbf{e}_j$ using terms of the form $c.\mathtt{Pr}(p(t))$ where p is an atomic proposition and c a rational constant. This means that using linear sums of probability expressions of the form $\mathtt{Pr}(p(t))$ for some proposition p, we can capture $\mathbf{f}^T M^t \mathbf{g}$ for arbitrary rational vectors $\mathbf{f}, \mathbf{g}$.

We show that we can find a stochastic matrix M such that for each function $f(t) = t \cdot \lambda^t$ and $f(t) = \lambda^t$ with $\lambda = -1/6$, there are rational vectors $\mathbf{f}, \mathbf{g}$ such that $f(t) = \mathbf{f}^T M^t \mathbf{g}$. Given a polynomial $p(x_1, \ldots, x_n)$, we can construct a variant polynomial q' over a larger set of variables, such that an appropriate substitution of such functions $t_i \cdot \lambda^{t_i}$ and λ^{t_i}, for the x_i and the additional variables yields an expression $\lambda^{k_1 t_1 + \ldots + k_n t_n} \cdot p(t_1, \ldots, t_n)$, where the k_i are constants. This has a zero in the $t_1 \ldots t_n$ iff $p(x_1, \ldots, x_n)$ has a zero. It follows using the result of the previous paragraph that mixed-time polynomial atomic probability formulas can express Hilbert's tenth problem. □

We remark that the possibility of encoding Hilbert's tenth problem is not immediate from the fact that we are dealing with polynomials, since our polynomials are over *rational* values generated in a very specific way from Markov chains, rather than arbitrary integers. Indeed, there are decidable logics containing polynomials, such as the theory of real closed fields [Tar51].

As noted in Section 2.5, formulas (allowed by Theorem 6) of the form $\mathtt{Pr}(\phi(t_1, \ldots, t_n)) \bowtie c$ can be written as a polynomial of probability expressions, so it is natural to ask whether such formulas also suffice to make the logic undecidable. This does not seem to be the case: the polynomials involved have only positive coefficients. Since Hilbert's tenth problem is trivially decidable for polynomials with only positive coefficients, our proof does not apply to this case.

5. CONCLUSION

Our results have by no means resolved Skolem's problem, which remains an apparent barrier to resolving the gap between the decidability results of [BRS06] and the undecidability results of the present paper.

However, in work to be presented elsewhere, we show that the results of [BRS06] can be extended both by reducing the set H_ϕ of cases that needs to be excluded to obtain decidability, as well as enhancing the expressiveness to cover epistemic probabilistic terms of the form $\mathtt{Pr}_i(\phi)$, interpreted with respect to the clock semantics.

Acknowledgment: The authors thank Igor Shparlinski and Min Sah for illuminating discussions on Skolem's problem, and Xiaowei Huang for helpful comments on a draft of the paper.

6. REFERENCES

[AAOW15] S. Akshay, T. Antonopoulos, J. Ouaknine, and J. Worrell. Reachability problems for Markov chains. *Information Processing Letters*, 115(2):155–158, 2015.

[ACD90] R. Alur, C. Courcoubetis, and D. L. Dill. Model-checking for real-time systems. In *Proc. of the Symposium on Logic in Computer Science*, pages 414–425, 1990.

[BRS06] D. Beauquier, A. M. Rabinovich, and A. Slissenko. A logic of probability with decidable model checking. *J. Log. Comput.*, 16(4):461–487, 2006.

[Buc60] J. R. Buchi. Weak second order arithmetic and finite automata. *Zeitscrift fur mathematische Logic und Grundlagen der Mathematik*, 6:66–92, 1960.

[CE81] E. M. Clarke and E. A. Emerson. The design and synthesis of synchronization skeletons using temporal logic. In *Proc. of the Workshop on Logics of Programs, IBM Watson Research Center, LNCS 131*, 1981.

[CES86] E. M. Clarke, E. A. Emerson, and A. P. Sistla. Automatic verification of finite-state concurrent systems using temporal logic specifications. *ACM Trans. Program. Lang. Syst.*, 8(2):244–263, 1986.

[CHHN14] J. Cassaigne, V. Halava, T. Harju, and F. Nicolas. Tighter undecidability bounds for matrix mortality, zero-in-the-corner problems, and more. *arXiv*, abs/1404.0644, 2014.

[CL04] A. Condon and R. J. Lipton. On the complexity of space bounded interactive proofs. In *Proc. of the Symp. on Foundations of Computer Science*, pages 462–467. IEEE, 2004.

[EGM07] K. Engelhardt, P. Gammie, and R. van der Meyden. Model checking knowledge and linear time: PSPACE cases. In *Proc. of the Int. Symp. on Logical Foundations of Computer Science LFCS*, pages 195–211, 2007.

[Eij04] D.J.N. Eijck. Dynamic epistemic modelling. *CWI. Software Engineering [SEN]*, (E 0424):1–112, 2004.

[ESPW03] G. Everest, I. Shparlinski, A. J. van der

Poorten, and T. Ward. *Recurrence sequences.* Amer. Math. Soc., 2003.

[FH94] R. Fagin and J. Y. Halpern. Reasoning about knowledge and probability. *J. ACM*, 41(2):340–367, 1994.

[FHM90] R. Fagin, J. Y. Halpern, and N. Megiddo. A logic for reasoning about probabilities. *Information and Computation*, 87(1/2):78–128, 1990.

[GM04] P. Gammie and R. van der Meyden. MCK: Model checking the logic of knowledge. In *Proc. Conf. on Computer-Aided Verification, CAV*, pages 479–483, 2004.

[Hal97] V. Halava. Decidable and undecidable problems in matrix theory. Technical Report 127, Turku Centre for Computer Science, University of Turku, Finland, 7 1997.

[Hal03] J. Y. Halpern. *Reasoning about Uncertainty.* MIT Press, Cambridge, MA, USA, 2003.

[HHHK05] V. Halava, T. Harju, M. Hirvensalo, and J. Karhumäki. Skolem's problem - on the border between decidability and undecidability. Technical Report 683, Turku Centre for Computer Science, University of Turku, Finland, 4 2005.

[Hir06] M. Hirvensalo. Improved undecidability results on the emptiness problem of probabilistic and quantum cut-point languages. Technical Report 769, Turku Centre for Computer Science, University of Turku, Finland, 5 2006.

[HLM11] X. Huang, C. Luo, and R. van der Meyden. Symbolic model checking of probabilistic knowledge. In *Proc. of the Conf. on Theoretical Aspects of Rationality and Knowledge*, pages 177–186, 2011.

[HM10] X. Huang and R. van der Meyden. The complexity of epistemic model checking: Clock semantics and branching time. In *Proc. ECAI 2010 - European Conf. on Artificial Intelligence*, pages 549–554, 2010.

[HSZ12] X. Huang, K. Su, and C. Zhang. Probabilistic alternating-time temporal logic of incomplete information and synchronous perfect recall. In *Proc. AAAI*, 2012.

[HV91] J. Y. Halpern and M. Y. Vardi. Model checking vs. theorem proving: A manifesto. In *Proc. of the Int. Conf. on Principles of Knowledge Representation and Reasoning*, pages 325–334, 1991.

[Kam68] H. W. Kamp. *Tense logic and the theory of linear order.* PhD thesis, University of California, Los Angeles, 1968.

[KNN$^+$08] M. Kacprzak, W. Nabiałek, A. Niewiadomski, W. Penczek, A. Półrola, M. Szreter, B. Woźna, and A. Zbrzezny. Verics 2007 - a model checker for knowledge and real-time. *Fundamenta Informaticae*, 85(1):313–328, 2008.

[KSK76] J. G. Kemeny, J. L. Snell, and A. W. Knapp. *Denumerable Markov Chains.* Springer-Verlag, 1976.

[LQR09] A. Lomuscio, H. Qu, and F. Raimondi. MCMAS: a model checker for the verification of multi-agent systems. In *Proc. Conf. on Computer-Aided Verification*, pages 682–688, 2009.

[Mat93] Yuri V. Matiyasevich. *Hilbert's Tenth Problem.* MIT Press, 1993.

[MNP08] O. Maler, D. Nickovic, and A. Pnueli. Checking temporal properties of discrete, timed and continuous behaviors. In *Pillars of Computer Science, Essays Dedicated to Boris (Boaz) Trakhtenbrot on the Occasion of His 85th Birthday*, pages 475–505, 2008.

[MS99] R. van der Meyden and N. V. Shilov. Model checking knowledge and time in systems with perfect recall. In *Proc. FST-TCS*, pages 432–445, 1999.

[MW03] R. van der Meyden and K-S. Wong. Complete axiomatizations for reasoning about knowledge and branching time. *Studia Logica*, 75(1):93–123, 2003.

[OW14] J. Ouaknine and J. Worrell. Positivity problems for low-order linear recurrence sequences. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pages 366–379, 2014.

[Paz71] A. Paz. *Introduction to probabilistic automata.* Academic Press, 1971.

[RKNP04] J. Rutten, M. Kwiatkowska, G. Norman, and D. Parker. *Mathematical Techniques for Analyzing Concurrent and Probabilistic Systems, P. Panangaden and F. van Breugel (eds.)*, volume 23 of *CRM Monograph Series.* American Mathematical Society, 2004.

[Sha49] C. Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28(4):656–715, 1949.

[Sko34] T. Skolem. Ein Verfahren zur Behandlung gewisser exponentialer Gleichungen und diophatischer Gleighungen. In *8de Skand. mat. Kongr. Forth, Stockholm*, 1934.

[Tar51] A. Tarski. *A Decision method for elementary algebra and geometry.* Univ. of California Press, 2nd edition, 1951.

[TMS84] R. Tijdeman, M. Mignotte, and T.N. Shorey. The distance between terms of an algebraic recurrence sequence. *Journal für die reine und angewandte Mathematik*, 349:63–76, 1984.

Relating Knowledge and Coordinated Action: The Knowledge of Preconditions Principle

Yoram Moses[*]
Technion—Israel Institute of Technology
moses@ee.technion.ac.il

ABSTRACT

The ***Knowledge of Preconditions*** principle (**K*o*P**) is proposed as a widely applicable connection between knowledge and action in multi-agent systems. Roughly speaking, it asserts that if some condition φ is a necessary condition for performing a given action α, then *knowing* φ is also a necessary condition for performing α. Since the specifications of tasks often involve necessary conditions for actions, the **K*o*P** shows that such specifications induce knowledge preconditions for the actions. Distributed protocols or multi-agent plans that satisfy the specifications must ensure that this knowledge be attained, and that it is detected by the agents as a condition for action. The knowledge of preconditions principle is formalised in the runs and systems framework, and is proven to hold in a wide class of settings. Well-known connections between knowledge and coordinated action are extended and shown to derive directly from the **K*o*P**: a *common knowledge of preconditions* principle is established showing that common knowledge is a necessary condition for performing simultaneous actions, and a *nested knowledge of preconditions* principle is proven, showing that coordinating actions to be performed in linear temporal order requires a corresponding form of nested knowledge.

Keywords

Knowledge, multi-agent systems, common knowledge, nested knowledge, coordinated action, knowledge of preconditions principle

1. INTRODUCTION

While epistemology, the study of knowledge, has been a topic of interest in philosophical circles for centuries and perhaps even millennia, in the last half century it has seen a flurry of activity and applications in other fields such as AI [19], game theory [2] and distributed computing [13]. At least in the latter two fields a particular, information-based, notion of knowledge plays a prominent and useful role.

This paper proposes an essential connection between knowledge and action in such a setting. Using $\texttt{does}_i(\alpha)$ to denote "*Agent i is performing action* α" and $K_i\varphi$ to denote that Agent i knows the fact φ, the connection can intuitively be formulated as follows:

The Knowledge of Preconditions Principle (**K*o*P**):

If	$\boldsymbol{\varphi}$	is a necessary condition for $\texttt{does}_i(\alpha)$
then	$\boldsymbol{K_i\varphi}$	is a necessary condition for $\texttt{does}_i(\alpha)$

This statement appears deceptively simple. In fact, many successful applications of knowledge to the design and analysis of distributed protocols over the last three decades are rooted in the **K*o*P**. Moreover, some of the deeper insights obtained by knowledge theory in this field can be derived in a fairly direct fashion from the **K*o*P**. We will argue and demonstrate that this principle lies at the heart of coordination in many distributed and multi-agent systems.

This paper is structured as follows. Section 1.1 illustrates the central role of knowledge in a natural distributed systems application. Section 1.2 provides a high-level discussion of the knowledge of preconditions principle and its connection to coordinating actions. In Section 2 we review and discuss the modelling of knowledge in the runs and systems model of distributed systems based on [11]. A formal statement and proof of the **K*o*P** are presented in Section 3. Then, in Section 4, the **K*o*P** is used to establish a *common knowledge of preconditions* principle. It states that in order to perform simultaneously coordinated actions, agents must first attain common knowledge of any of the actions' preconditions. An example of its use is provided in Section 4.1. Section 5 present an additional use of the **K*o*P**, and shows that coordinating a sequence of actions to occur in a prescribed temporal order requires attaining nested knowledge of their preconditions. Finally, Section 6 discusses additional applications, extensions and future directions.

1.1 The Case for Knowledge in Distributed Systems

Why should knowledge play a central role in distributed computing? As pointed out in [13], most everyone who designs or even just tries to study the workings of a distributed protocol is quickly found talking in terms of knowledge, making statements such as *"once the process receives an acknowledgement, it knows that the other process is ready…"*. An essential aspect of distributed systems is the fact that an agent chooses which action to perform based on the local information available to it, which typically provides only a partial view of the overall state of the system. To get a sense of the role of knowledge in distributed systems, consider the following example.

Example 1. *Given is a distributed network modeled by a*

[*] The Israel Pollak academic chair at Technion. This work was supported in part by ISF grant 1520/11.

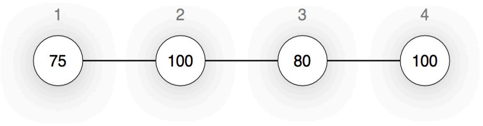


Figure 1: A simple four-agent system

*graph, with agents located at the nodes, and the edges standing for communication channels (see Figure 1). In the problem we shall call **Computing the Max** (or* CTM *for short), each agent i starts out with a natural number* $v_i \in \mathbb{N}$ *as an initial value. The goal is to have Agent 1 print the maximum of all of the initial values (we denote this value by* Max*), and print nothing else. In the instance depicted in Figure 1, the maximal value happens to be 100. Initially, Agent 1 clearly can't print its own initial value of 75. Suppose that Agent 1 receives a message* $\mu \triangleq$ *"*$v_2 = 100$*" from Agent 2 reporting that its value is* 100*. At this point, Agent 1 has access to the maximum, and printing 100 would satisfy the problem specification. Compare this with a setting that is the same in all respects, except that Agent 3's value is* $v_3 = 150$*. In this case, of course,* $\mathsf{Max} \neq 100$ *and so printing 100 is forbidden. But if Agent 1 can receive the same message* μ *under similar circumstances in both scenarios, then it is unable to distinguish whether or not* $\mathsf{Max} = 100$ *upon receiving* μ*. Intuitively, even in the first scenario, the agent does not* know *that* $\mathsf{Max} = 100$*.*

What information does Agent 1 need, then, in order to be able to print the maximum? Notice that it, is not necessary, in general, to collect all *of the initial values in order to print the maximum. For example, suppose that the agents follow a bottom-up protocol in which values are sent from right to left, starting from Agent 4, and every agent passes to the left the larger of its own value and the value it received from its neighbor on the right (if such a neighbor exists). In this protocol, Agent 1 can clearly print the maximum after receiving the message "*$v_2 = 100$*", and seeing just one value besides its own. Interestingly, even collecting all of the values is* not a sufficient condition *for printing the* Max*. Imagine a setting in which the network is as in Figure 1, but Agent 1 considers it possible that there are more than four nodes in the network. In this case, even if Agent 1 receives all (four) values, it may still need to wait for proof that there is no additional, larger, value in the system.* □

CTM is a simplified example in the spirit of many distributed systems applications. Leader election, for example, is often solved by computing a node with maximal ID [1, 17]. The solution to such a problem is typically in the form of a set of short computer programs (jointly constituting a *distributed protocol*) each executed at one of the nodes. When the nodes follow such a protocol, the resulting execution should satisfy the problem specification. Of course, the programs are written in a standard programming language, without any reference to knowledge or possibility. In the vast majority of cases, the programs in question do not enumerate and/or explore possible states or scenarios. Indeed, the program designer is typically unfamiliar with formal notions of knowledge. This being the case, what sense does it make to talk of Agent 1 in Example 1 "knowing" or "not knowing" that $\mathsf{Max} = c$? Can it make sense to say that the Agent "considers it possible that there may be more than four nodes in the system"? After all, we may be talking about a ten-line program. It has no soul. Does it have thoughts, doubts and mental states?

Since agents act based on their local information, a protocol designer must ensure that agents obtain the necessary information for a given task, and that this information is applied correctly. Using the information-based notion of knowledge, the designer can ascribe knowledge to an agent without requiring it to have a soul, feelings, and self-awareness. As seen in the CTM example, it is natural to think in terms of whether or not Agent 1 knows $\mathsf{Max} = c$ at any given point in a run of CTM. (A formal definition of knowledge will be provided in Section 2.) Suppose that a protocol is designed to solve CTM in networks that may have a variety of sizes. If Agent 1 does not start out with local information ensuring that there are no more than four nodes in the system, then from the point of view of an outside observer the agent can be thought of as "considering it possible" that there may be more than four nodes.

Even in a simple network as in Figure 1, the CTM problem can be posed in different models, which can differ in essential aspects. A solution to CTM in one model might not solve the problem in another model. Indeed, the rationale behind distinct solutions, as well as their details, may vary considerably. Are there common features shared by all solutions to CTM?

Interestingly, all solutions to CTM, in all models, share one property: Agent 1 must ***know*** that $\boldsymbol{\mathsf{Max} = c}$ in order to print the value c. Indeed, the ability to print the answer in a protocol for CTM reduces to detecting when the Max value is known. Of course, once Agent 1 knows that $\mathsf{Max} = c$ it can safely print c. Hence, knowing that $\mathsf{Max} = c$ is not just necessary, but also a sufficient condition for printing c. The CTM problem shows that knowledge and attaining knowledge can be a central and crucial aspect of a standard distributed application.

The need to know $\mathsf{Max} = c$ in solving CTM suggests that we consider a natural question: *When does Agent 1 know that* $\mathsf{Max} = c$*?* The answer is less straightforward than we might initially expect. What is known depends in a crucial way on the protocol that the agents are following. Thus, in the setting of Example 1, if the agents follow the bottom-up protocol, then Agent 1 knows the maximum once it receives a single message from Agent 2. Knowledge is also significantly affected by features of the model. In CTM, if there is an upper bound (say 100) on the possible initial values, then an agent that sees this value knows the maximum. Knowledge about the network topology and properties of communication play a role as well. For example, consider a model in which Agent 1 has a clock, and a single clock cycle suffices for a message to be delivered, digested, and acted upon. Suppose that the protocol is such that all agents start simultaneously at time 0 and an agent forwards a value towards Agent 1 only if this value is larger than any value it has previously sent. Then in the network of Figure 1 Agent 1 will receive a message with value 100 from Agent 2 at time 1, and no further messages. If Agent 1 knows that the diameter of the network is 3, it will not know the maximum upon receiving this message. However, without receiving any further messages, at time 3 Agent 1 *will* know that the maximum is 100; no larger value can be lurking in the system.

1.2 KoP and Coordination

The fact that $\mathsf{Max} = c$ is a necessary condition for printing c is an essential feature of the CtM problem. We have argued that, in fact, $K_1(\mathsf{Max} = c)$ is also a necessary condition for printing c, as the **K*o*P** would suggest. But this is just one instance. Let us briefly consider another example.

Example 2. *Consider a bank whose ATMs are designed in such a way that an ATM will dispense cash only to a customer whose account shows a sufficiently large positive balance. Along comes Alice, who has a large positive balance, and tries to obtain a modest sum from the ATM. On this day, however, the ATM is unable to communicate with the rest of the bank and it declines to pay Alice. Thus, despite the fact that Alice has good credit, the ATM frustrates her and denies her request. Apparently, given its specification, the ATM has no choice. Intuitively, in order to satisfy the credit restriction, the ATM needs to* know *that a customer has good credit before dispensing cash. If the ATM may pay a customer that is not known to have good credit, there will be possible scenarios in which the ATM will violate its specification, and pay a customer that does not have credit. Notice, however, that the specification said nothing about the ATM's knowledge. It only imposed a restriction on the ATM's action, based on the state of Alice's account.* □

Both the CtM problem and the ATM example are instances in which the **K*o*P** clearly applies. The intuitive argument for why the **K*o*P** should apply very broadly is straightforward. If φ is a necessary condition for performing α, and agent i ever performs α without knowing φ, then there should be a possible scenario that is indistinguishable to agent i, in which φ does not hold. Since the two scenarios are indistinguishable, the agent can perform α in the second scenario, and violate the requirement that φ is a necessary condition. A formal statement and proof requires a definition of necessary conditions, knowledge, as well as capturing a sense in which an action at one point implies the same action at any other, indistinguishable, point. This will be done in Section 3.

Most tasks in distributed systems are described by way of a specification. Such specifications typically impose a variety of necessary conditions for actions. The **K*o*P** implies that even though such specifications often do not explicitly discuss the agents' knowledge, they do in fact impose knowledge preconditions. Observe that the **K*o*P** applies to a task regardless of the means that are used to implement it. Any engineer implementing a particular task will have to ensure that preconditions are known when actions are taken. This is true whether or not the engineer reasons explicitly in terms of knowledge, and it is true even if the engineer is not even aware of the knowledge terminology. (Normally, neither may be the case, of course.) The need to satisfy the **K*o*P** suggests that the design of distributed implementations must involve at least two steps. One is to make sure that the required knowledge is made available to an agent who needs to performed a prescribed action, and the other is ensuring that the agent detect that it knows the required preconditions. This is quite different from common practice in engineering distributed implementations [28].

We remark that the **K*o*P** can be expected to hold in a variety of multi-agent settings well beyond the realm of distributed systems. Thus, for example, suppose that a jellyfish is designed so that it will never sting its own flesh. By the **K*o*P**, the cell activating the sting at a given point needs to *know* that it is not stinging the jellyfish's body when it "fires" its sting. The jellyfish is thus designed with some form of a "friend or foe" mechanism that is used in the course of activating the sting. Various biological activities can similarly be considered in light of the **K*o*P**: How does the organism know that certain preconditions are met? Our last example will come from the social science arena. Suppose that a society designs a legal system, that is required to satisfy the constraint that only people who are guilty of a particular crime are ever put in jail based on this crime. By the **K*o*P**, the judge (or jury) must *know* that the person committed this crime in order to send him to jail.

As discussed above, specifications impose preconditions. Typically, these conditions relate an action to facts about the world (e.g., the maximal value, or the customer's good credit). In many cases, however, actions of different agents need to be coordinated. Consider a variant of CtM in which in addition to Agent 1 printing the maximum, Agent 4 needs to perform an action (say print the same value or print the minimal value), but not before Agent 1 does. Then Agent 1 performing her action is a condition for 4's action. In particular, Agent 4 would need to know that Agent 1 has already come to know $\mathsf{Max} = c$ for some c before 4 acts. In some cases, the identity of actions performed needs to be coordinated.

For a final example, suppose that Alice 1 should perform an action α_A only if Bob performs an action α_B at least 5 time steps earlier. Then she needs to know that Bob acted at least 5 steps before when she acts. Indeed, if ψ is a necessary condition for α', then Alice must know that Bob performed "Bob knew ψ at least 5 time steps ago" when she acts (see [4, 5]). As these examples illustrate, given **K*o*P**, coordination can give rise to nested knowledge.

Simple instances of the **K*o*P** are often quite straightforward. Ensuring and detecting $K_1(Max = c)$ is often fairly intuitive, and it not justify the overhead involved in developing a theory of knowledge for multi-agent systems. However, satisfying statements involving nested knowledge in particular models of computation can quickly become nontrivial. For this, it is best to have a clear mathematical model of knowledge in multi-agent systems. The next section reviews the runs and systems model.

2. MODELING KNOWLEDGE USING RUNS AND SYSTEMS

We now review the runs and systems model of knowledge of [11, 13]. The interested reader should consult [11] for more details. A ***global state*** is an "instantaneous snapshot" of the system at a given time. Let $\mathcal{G}$ denote a set of global states. Time will be identified with the natural numbers $\mathbb{N} = \{0, 1, 2, \ldots\}$ for ease of exposition. A ***run*** is a function $\boldsymbol{r}\colon \mathbb{N} \to \mathcal{G}$ associating a global state with each instant of time. Thus, $r(0)$ is the run's initial state, $r(1)$ is the next global state, and so on. A ***system*** is a set $\boldsymbol{R}$ of runs. The same global state can appear in different runs, and in some systems may even appear more than once in the same run.

A central notion in our framework is that of an agent's *local state*, whose role is to capture the agent's local information at a given point. The precise details of the local state depend on the application. It could be the complete contents of an agent's memory at the given instant, or the complete sequence of events that it has observed so far. for example. The rule of thumb is that the local state should consist of the local information that the agent may use when deciding which actions to take. Thus, for example, if agents are finite-state machines, it is often natural to identify an agent's local state with the automaton state that it is in. Formally, we assume that every global state determines a unique ***local state*** for each agent. We denote agent i's local state in the global state $r(t)$ by $r_i(t)$. Moreover, a global state with n agents $\mathbf{A} = \{1, \ldots, n\}$ will have the form $r(t) = \langle r_e(t), r_1(t), \ldots, r_n(t)\rangle$, where $r_e(t)$ is called the local state of the *environment*, and will serve to represent all aspects of the global state that are not included in the agents' local states. For example, it could represent messages in transit, the current topology of the network including what links may be down, etc.

2.1 Syntax and Semantics

We are interested in a propositional logic of knowledge, in which propositional facts and epistemic facts can be expressed. Facts will be considered to be true or false at a point (r,t), with respect to a system R. More formally, given a set Φ of primitive propositions and a set $\mathbb{P} = \{1, \ldots, n\}$ of the agents in the system, we define a propositional language $\mathcal{L}_n^K(\Phi)$ by closing Φ under negation '$\neg$' and conjunction '$\wedge$', as well as under knowledge operators K_i for all $i \in \mathbb{P}$ (see [14]). Thus, for example, if $p, q \in \Phi$ are primitive propositions and $i, j \in \mathbb{P}$ are agents, then $\neg K_i p \wedge K_j K_i \neg K_j q$ is a formula in $\mathcal{L}_n^K(\Phi)$. We typically omit the set Φ and call $\mathcal{L}_n^K$ the language for knowledge with n agents.

In a multi-agent system facts about the world, as well as the knowledge that agents have, can change dynamically from one time point to the next. We thus consider the truth of formulas of $\mathcal{L}_n^K$ at *points* of a system R, where a point is a pair $\boldsymbol{(r,t)} \in R \times \mathbb{N}$, and it is used to refer to time t in the run r. We denote the set of points of a system R by $\mathsf{Pts}(R) \triangleq R \times \mathbb{N}$. Points will play the role of states of a Kripke structure.

The set Φ of primitive propositions used in the analysis of any given multi-agent system R will depend on the application. Their truth at the points of the system needs to be explicitly defined. This is done by an *interpretation* $\pi : \Phi \times \mathsf{Pts}(R) \to \{T, F\}$, where $\pi\big(q, (r,t)\big) = T$ means that the proposition q holds at (r,t). Formally, an *interpreted system* w.r.t. a set Φ of primitive propositions is a pair (R, π) consisting of the system R and interpretation π for Φ over $\mathsf{Pts}(R)$. Just as we typically omit explicit reference to Φ, we shall omit π as well, when this is unambiguous.

We assume from here on that the environment's state $r_e(t)$ in a global state $r(t)$ contains a "*history*" component h that records all actions taken by all agents at times 0,1,...,$t-1$. Formally, we take h to be a set of triples $\langle \alpha, i, t'\rangle$, which grows monotonically in time. An action α is considered to be performed by i at the point (r,t) if and only if the triple $\langle \alpha, i, t\rangle$, denoting that action α was performed by agent i at time t, appears in the history component h of $r_e(t')$ for all times $t' > t$.[1] For the analysis in this paper, we will also assume that Φ includes propositions of the form $\mathtt{does}_i(\alpha)$ and $\mathtt{did}_i(\alpha)$ for agents $i \in \mathbb{P}$ and actions α. With this assumption, what actions are performed at any given point (r,t) is uniquely determined by the run r.

We will consider interpretations π that, on these propositions, are defined by

$$\pi\big(\mathtt{does}_i(\alpha), (r,t)\big) = T \quad \text{iff} \quad \text{agent } i \text{ performs } \alpha \text{ at } (r,t)$$

$$\pi\big(\mathtt{did}_i(\alpha), (r,t)\big) = T \quad \text{iff} \quad \pi\big(\mathtt{does}_i(\alpha), (r,t')\big) = T \text{ holds for some } t' \le t$$

We allow $t' = t$ in the definition of $\mathtt{did}_i(\alpha)$ for technical convenience; it simplifies our later analysis slightly.

Our model of knowledge will follow the standard Kripke-style possible worlds approach. The possibility relations that we use are induced directly from the system R being analyzed; two points are considered indistinguishable to an agent if its local states at the two points are the same. More formally:

Definition 2.1. *If $r_i(t) = r'_i(t')$, then (r,t) and (r',t') are called **indistinguishable to i**, denoted by $(r,t) \approx_{\boldsymbol{i}} (r',t')$.*

Formulae of $\mathcal{L}_n^K$ are interpreted at a point (r,t) of an interpreted system (R,π) by means of the satisfaction relation '$\models$', which is defined inductively by:

$(R,r,t) \models p$ iff $(r,t) \in \pi(p)$;

$(R,r,t) \models \neg\varphi$ iff $(R,r,t) \not\models \varphi$;

$(R,r,t) \models \varphi \wedge \psi$ iff both $(R,r,t) \models \varphi$ and $(R,r,t) \models \psi$;

$(R,r,t) \models K_i\varphi$ iff $(R,r',t') \models \varphi$ for all $(r',t') \in \mathsf{Pts}(R)$ such that $(r',t') \approx_{\boldsymbol{i}} (r,t)$.

We say that $\boldsymbol{\varphi}$ is ***valid in*** the system $\boldsymbol{R}$, and write $\boldsymbol{R} \models \boldsymbol{\varphi}$, if $(R,r,t) \models \varphi$ for all points $(r,t) \in \mathsf{Pts}(R)$. We say that $\boldsymbol{\varphi}$ ***validly implies*** $\boldsymbol{\psi}$ ***in*** $\boldsymbol{R}$ if $\varphi \Rightarrow \psi$ is valid in R. Since, by Definition 2.1, the $\approx_i$ relations are equivalence relations, each knowledge operator K_i satisfies the S5 axiom system [14]. In particular, it satisfies the ***knowledge property*** (or axiom) that $K_i\varphi \Rightarrow \varphi$ is valid in all systems.

It is instructive to relate our modeling using runs and systems to standard multi-agent Kripke structures. As shown in [11], for every system R there is a corresponding Kripke structure $M_R = (S_R, \pi, \sim_1, \ldots, \sim_n)$ for n agents such that $S_R = \mathsf{Pts}(R)$ and '$\sim_i$' = '$\approx_i$' for every i. For every point $(r,t) \in \mathsf{Pts}(R) = S_R$ and formula $\varphi \in \mathcal{L}_n^K(\Phi)$ it is the case that $(R,r,t) \models \varphi$ iff $M_R, (r,t) \models \varphi$.

The system R will determine the space of possible runs and possible points, which play a crucial role in determining the truth of facts involving knowledge. For example, consider a run r in which Alice sends Bob a message at time 1, and Bob receives it at time 2. If R is a system in which messages may be lost, or may take longer than one time step to be delivered, then Alice would not know at time 2 (i.e., w.r.t. $(R,r,2)$) that her message has been delivered, because there is another run $r' \in R$ that she cannot tell apart from r at time 2, in which her message is not (or not yet) delivered by that time. The same run r also belongs to another system R' in which messages are always reliably delivered in

[1] Our definition does not imply or assume that the actions are observed, observable or recorded by any of the agents. Whether that may be the case depends on the application.

exactly one round. With respect to $(R', r, 2)$, however, Alice *would* know at time 2 that her message has been delivered.

Our definition of knowledge is rather flexible and widely applicable. The set R of the possible runs immediately induces what the agents know. Observe that the definition of knowledge is completely external. It ascribes knowledge to agents in the system even if the protocol they follow, as well as the actions that they perform, do not involve the knowledge terminology in any way. Moreover, the agents do not need to be complex or sophisticated for the definition to apply. Indeed, in a model of a very simple system consisting of a bed lamp and its electric cable, a switch in the OFF state can be said to know that the lamp is not lit; what the same switch would know in the ON state would depend on the system R under consideration, which determines the runs considered possible. E.g., if R contains a run in which the lamp is burnt out, then in the ON state the switch would not know that the lamp is shining light. On the other hand, if the lamp can never burn out, and the cord, plug and switch are in proper working order in all runs of R, then in the ON state the switch *would* know that the lamp is shining light. As this example shows, knowledge under this definition does not require the "knower" to compute what it knows. Indeed, this definition of knowledge is not sensitive to the computational complexity of determining what is known. In most cases, of course, we will ascribe knowledge to agents or components that can perform actions, which is not the case in the light switch example. And agents might need to explicitly establish whether they know relevant facts. We now provide a statement and proof of the knowledge of preconditions principle **K*o*P**.

3. FORMALIZING THE KNOWLEDGE OF PRECONDITIONS PRINCIPLE

Intuitively, the **K*o*P** states that if a particular fact ψ is a necessary condition for an agent to perform an action α, then the agent must in fact *know* ψ in order to act. In other words, *knowing* ψ is also a necessary condition for performing the action. We formalize the claim and prove it as follows. We say that ***ψ is a necessary condition for*** $\mathtt{does}_i(\alpha)$ ***in R*** if $(R, r, t) \models \mathtt{does}_i(\alpha)$ holds only if $(R, r, t) \models \psi$, for all $(r, t) \in \mathsf{Pts}(R)$. Clearly, the customer's good credit is a necessary condition for the ATM dispensing cash. That is, suppose that a bank makes use of a correct implementation of an ATM protocol, which satisfies the credit requirement. Then, in the system R consisting of the set of all possible histories (runs) of the bank's (and the ATM's) transactions, good credit is a necessary condition for receiving cash from the ATM.

It is often of interest to consider facts whose truth depends only on a given agent's local state. Such, for example, may be the receipt of a message, or the observation of a signal, by the agent. Whether $x = 0$ for a local variable x, for example, would be a natural local fact. Moreover, if an agent has perfect recall, then any events that it has observed in the past will give rise to local facts. Finally, since knowledge is defined based on an agent's local state, then a fact of the form $K_i\varphi$ constitutes a local fact. Indeed, there is a simple way to define the local facts above, using knowledge. Namely, we say that ***φ is i-local in R*** if $R \models (\varphi \Rightarrow K_i\varphi)$.

The formalism of [11] defines protocols as explicit objects, and defines *contexts* that describe the possible initial states and the model of computation. This provides a convenient and modular way of constructing systems. Namely, given a protocol P and a *context* γ, the system $R = R(P, \gamma)$ is defined to be the set of all runs of protocol P in γ. The runs of this system embody all of the properties of the context, as they arise in runs of P. This includes, for example, any timing assumptions, possible values encountered, possible topologies of the network, etc. They also embody the relevant properties of the protocol, because in all runs considered possible the agents follow P.

In this paper, we do not define protocols and contexts. Rather, we treat the **K*o*P** in a slightly simpler and more abstract setting. We say that an action α is a ***conscious action for i in R*** if $\mathtt{does}_i(\alpha)$ is an i-local fact in R, so that whenever $(R, r, t) \models \mathtt{does}_i(\alpha)$ holds, $(R, r, t) \models K_i\mathtt{does}_i(\alpha)$ holds as well. In other words, the fact that α is a conscious action for i in R implies that if α is ever performed at a point of R in which i's local state is ℓ_i, then α must be performed whenever i's state is ℓ_i. Conscious actions are quite prevalent in many systems of interest. For example, suppose that agent i follows a deterministic protocol, so that its action at any given point is a function of its local state. If, in addition, agent i is allowed to move at every time step, then all of its actions are conscious actions.

We are now ready to prove a formal version of the **K*o*P**:

THEOREM 3.1 (THE **K*o*P** PRINCIPLE). *Let α be a conscious action for i in R. If ψ is a necessary condition for* $\mathtt{does}_i(\alpha)$ *in R, then $K_i\psi$ is also a necessary condition for* $\mathtt{does}_i(\alpha)$ *in R.*

PROOF. We will show the contrapositive. Let α be a conscious action for i in R, and assume that $K_i\psi$ is not a necessary condition for $\mathtt{does}_i(\alpha)$ in R. Namely, there exists a point $(r, t) \in \mathsf{Pts}(R)$ such that both $(R, r, t) \models \mathtt{does}_i(\alpha)$ and $(R, r, t) \not\models K_i\psi$. Given the latter, we have by the definition of '$\models$' for K_i that there exists a point $(r', t') \in \mathsf{Pts}(R)$ such that both $(r', t') \approx_i (r, t)$ and $(R, r', t') \not\models \psi$. Since α is a conscious action for i in R and $(R, r, t) \models \mathtt{does}_i(\alpha)$ we have that $(R, r, t) \models K_i\mathtt{does}_i(\alpha)$. It follows from $(r', t') \approx_i (r, t)$ by the definition of '$\models$' for K_i that $(R, r', t') \models \mathtt{does}_i(\alpha)$ holds. But since $(R, r', t') \not\models \psi$, we conclude that ψ is not a necessary condition for $\mathtt{does}_i(\alpha)$ in R, establishing the countrapositive claim. □

Theorem 3.1 applies to all multi-agent systems. It immediately implies, for example, that $K_{atm}(\mathtt{good\_credit})$ is a necessary condition for dispensing cash. The theorem is model independent; it does not depend on timing assumptions, on the topology of the system (even on whether agents communicate by message passing or via reading and writing to registers in a shared memory), or on the nature of the activity that is carried out. For every necessary condition for a conscious action, *knowing* that the condition holds is also a necessary condition.

4. COORDINATING SIMULTANEOUS ACTIONS

Recall that the language $\mathcal{L}_n^K$ contains formulas in which knowledge operators can be *nested* to arbitrary finite depth. It is sometimes useful to consider a state of knowledge called *common knowledge* that goes beyond any particular nested formula. Intuitively, a fact ψ is common knowledge if everyone knowing that everyone knows . . . , that everyone knows the fact ψ, to every finite depth. Common knowledge has a number of equivalent definitions, one of which is as follows:

Definition 4.1 (Common Knowledge). *Fix a set of agents G and a fact ψ. We denote by $C_G\psi$ the fact that* ***ψ is common knowledge to G****. Its truth at points of a system R is defined by:*

$$(R,r,t) \models C_G\psi \quad \textit{iff} \quad (R,r,t) \models K_{i_1}K_{i_2}\cdots K_{i_m}\psi \text{ for all } \langle i_1,i_2,\ldots,i_m\rangle \in G^m, \text{ and all } m \geq 1.$$

Common knowledge, a term coined by Lewis in [18], plays an important role in the analysis of games [2], distributed systems [13], and many other multi-agent settings. Clearly, common knowledge is much stronger than "plain" knowledge. Indeed, $C_G\psi$ validly implies $K_j\psi$, for all agents $j \in G$. Since common knowledge requires infinitely many facts to hold, it is not *a priori* obvious that $C_G\varphi$ can be attained at a reasonable cost, or even whether it can ever be attained at all, in settings of interest (see [7, 11, 13]). We will now show that there are natural applications for which attaining common knowledge is essential.

Intuitively, distinct actions are simultaneous in R if they can only be performed together; whenever one is performed, all of them are performed simultaneously. It is possible to define simultaneous coordination formally in terms of necessary conditions:

Definition 4.2 (Simultaneous Actions). *Let G be a set of agents. We say that a set of actions $\boldsymbol{A} = \{\alpha_i\}_{i\in G}$* ***is*** *(necessarily)* ***simultaneous in R*** *if $\texttt{does}_i(\alpha_i)$ is a necessary condition for $\texttt{does}_j(\alpha_j)$ in R, for all $i, j \in G$.*

Suppose that the actions in A are simultaneous in R in the above sense. Then the **K*o*P** immediately implies (by Theorem 3.1) that a necessary condition for performing an action in A is knowing that the other actions are also (currently) being performed. In fact, however, much more must be true. We now present a strong variant of the **K*o*P**, which shows that in order to perform simultaneous actions agents must attain common knowledge of their necessary conditions. Notice that in order to allow a set of actions by the agents in G to be simultaneous, the system R must be sufficiently deterministic to ensure that if $i, j \in G$ are distinct agents and $(R,r,t) \models \texttt{does}_i(\alpha)$ holds, then j will be scheduled to perform an action at (r,t). For otherwise, there would be no way to ensure simultaneous execution of the actions by the agents in G. Conscious actions fit this setting well in this case. We proceed as follows.

Theorem 4.3 (C-K of Preconditions). *Let G be a set of agents and let $A = \{\alpha_i\}_{i\in G}$ be a set of necessarily simultaneous actions in the system R. Moreover, suppose that each action $\alpha_i \in A$ is a conscious action for its agent i in R. If ψ is a necessary condition for $\texttt{does}_i(\alpha_i)$ for some $i \in G$, then $C_G\psi$ is a necessary condition for $\texttt{does}_j(\alpha_j)$, for all $j \in G$.*

Proof. Assume that A is a set of necessarily simultaneous actions for G in R. It is straightforward to show the following claim

Observation 1. *Let $\alpha_i, \alpha_j \in A$ be the actions for agents i and j, respectively. If a fact φ is a necessary condition for $\texttt{does}_i(\alpha_i)$ in R then φ is also a necessary condition for $\texttt{does}_j(\alpha_j)$ in R.*

To prove this observation notice that, by assumption, both (a) $R \models \texttt{does}_j(\alpha_j) \Rightarrow \texttt{does}_i(\alpha_i)$ and (b) $R \models \texttt{does}_i(\alpha_i) \Rightarrow \varphi$ hold. For all $(r,t) \in \mathsf{Pts}(R)$, if $(R,r,t) \models \texttt{does}_j(\alpha_j)$ then $(R,r,t) \models \texttt{does}_i(\alpha_i)$ by (a) and so $(R,r,t) \models \varphi$ by (b). Thus, φ is a necessary condition for $\texttt{does}_j(\alpha_j)$ in R.

Assume that ψ is a necessary condition for $\texttt{does}_i(\alpha_i)$, for some $i \in G$. We shall prove by induction on $m \geq 0$ that $K_{i_1}K_{i_2}\cdots K_{i_m}\psi$ is a necessary condition for $\texttt{does}_j(\alpha_j)$ in R, for every $j \in G$ and *all* sequences $\langle i_1,\ldots,i_m\rangle \in G^m$ (of m agent names from G). This will establish that $(R,r,t) \models \texttt{does}_j(\alpha_j)$ implies $(R,r,t) \models C_G\psi$ for all $(r,t) \in \mathsf{Pts}(R)$, and thus $C_G\psi$ is a necessary condition for $\texttt{does}_j(\alpha_j)$ for all $j \in G$, as claimed.

- Base case: Let $m = 0$. The claim in this case is that if ψ is a necessary condition for $\texttt{does}_i(\alpha_i)$ then ψ is also a necessary condition for $\texttt{does}_j(\alpha_j)$. This is precisely Observation 1, with $\varphi \triangleq \psi$.

- Inductive step: Let $m \geq 1$, and assume that the claim holds for all $j' \in G$ and all sequences in G^{m-1}. Fix $j \in G$ and a sequence $\langle i_1,i_2,\ldots,i_m\rangle \in G^m$. Its suffix $\langle i_2,\ldots,i_m\rangle$ is a sequence in G^{m-1}. Thus, $K_{i_2}\cdots K_{i_m}\psi$ is a necessary condition for $\texttt{does}_{i_1}(\alpha_{i_1})$ by the inductive hypothesis for $m-1$ (applied to G^{m-1} and agent $j' = i_1 \in G$). Given that α_{i_1} is a conscious action by i_1, we can apply Theorem 3.1 to the necessary condition $K_{i_2}\cdots K_{i_m}\psi$ and obtain that $K_{i_1}K_{i_2}\cdots K_{i_m}\psi$ is a necessary condition for $\texttt{does}_{i_1}(\alpha_{i_1})$. By Observation 1 we have that $K_{i_1}K_{i_2}\cdots K_{i_m}\psi$ is also a necessary condition for $\texttt{does}_j(\alpha_j)$ in R, and we are done.

□

4.1 Common Knowledge and the Firing Squad Problem

As an illustration of the applicability of Theorem 4.3 to a concrete application, consider a simple version of the *Firing Squad* problem. In this instance, the set of agents G in the system must simultaneously perform an action (say each agent $i \in G$ should perform the action fire_i) in response to the receipt, by any agent in G, of a particular external input called a 'go' message. The fire_i action can stand for a simultaneous change in shared copies of a database, a public announcement at different sites of the system, or any other actions that need to take place simultaneously. Moreover, fire_i actions are allowed only if they are preceded by such a go message. For simplicity, we consider a case in which none of the agents in G may fail, and they all must satisfy the specification.

Let ψ_{go} be a proposition that is true at $(r,t) \in \mathsf{Pts}(R)$ if a go message is received by any of the agents in G at a point (r,t') of r at a time $t' \leq t$. According to the specification of the Firing Squad problem, ψ_{go} is a necessary condition for the fire_i actions. An immediate consequence of Theorem 4.3 is:

Corollary 4.4. *$C_G\psi_{\mathsf{go}}$ is a necessary condition for all fire_i actions in the Firing Squad problem.*

Given Corollary 4.4, any solution to the firing squad problem must first attain common knowledge that a go message has been received. It is well-known (see [10, 13]) that common knowledge of a fact is observed simultaneously at all agents it involves. Suppose that every $i \in G$ performs fire_i when

$C_G\psi_{\mathtt{go}}$ first holds. Since all agents in G will come to know that $C_G\psi_{\mathtt{go}}$ immediately, they will fire simultaneously, as required by the problem specification. Indeed, Theorem 4.3 shows that this is the first time at which they *can* perform according to a correct protocol. Implementing simultaneous tasks such as the Firing Squad therefore inherently involves, and often reduces to, ensuring and detecting $C_G\psi_{\mathtt{go}}$. Recall that depending on the properties of the system, attaining such common knowledge might be impossible in some cases, or it might incur a substantial cost in others. Just as in the case of the **K*o*P**, this necessity is not due to our formalism. It is only exposed by our analysis. In every protocol that implements such a task correctly, the firing actions cannot be performed unless $C_G\psi_{\mathtt{go}}$ is attained.

There is an extensive literature on using common knowledge to obtain optimal protocols for simultaneous tasks [8, 9, 16, 20, 21, 23, 24, 25, 26]. Typically, they involve an explicit proof that common knowledge of a particular fact is a necessary condition for performing a set A of necessarily simultaneous actions. Theorem 4.3 or a variant of it suited for fault-tolerant systems can be used to establish this result in all of these cases. Moreover, one of the main insights from the analysis of [13] and of [10] is that when simultaneous actions are performed, the participating agents have common knowledge that they are being performed. Theorem 4.3 is a strict generalization of this fact.

5. TEMPORALLY ORDERING ACTIONS

So far, we have seen two essential connections between knowledge and coordinated action: performing actions requires knowledge of their necessary conditions, and performing simultaneous actions requires common knowledge of their necessary conditions. We now further extend the connection between states of knowledge and coordination, by showing that temporally ordering actions depends on attaining nested knowledge of necessary conditions. Following [5], we define temporally ordered actions:

Definition 5.1 (Ben Zvi and Moses). *A sequence of actions* $\langle\boldsymbol{\alpha_1,\ldots,\alpha_k}\rangle$ *(for agents* $1,\ldots,k$*, respectively)* ***is*** *(linearly)* ***ordered in R*** *if* $\mathtt{did}_{j-1}(\alpha_{j-1})$ *is a necessary condition for* $\mathtt{does}_j(\alpha_j)$ *in R.*

Observe that this definition does not force an action α_j to occur in a run in which α_{j-1} occurs. Rather, if an action α_j is performed in a given run, then it must be preceded by all actions $\alpha_1,\ldots,\alpha_{j-1}$. Moreover, if we denote the time at which an action α_i is performed in a run r by t_i, then we require that $t_{j-1}\le t_j$ for every action α_j performed in r.

Claim 1. *Assume that the sequence* $\langle\alpha_1,\ldots,\alpha_k\rangle$ *is ordered in R. Then* $R\models\big(\mathtt{did}_j(\alpha_j)\Rightarrow\mathtt{did}_{j-1}(\alpha_{j-1})\big)$ *for all* $2\le j\le k$*.*

Proof. Assume that $(R,r,t)\models\mathtt{did}_j(\alpha_j)$. Then, by definition of $\mathtt{did}_j(\alpha_j)$, $(R,r,\hat{t})\models\mathtt{does}_j(\alpha_j)$ for some $\hat{t}\le t$. Since $\langle\alpha_1,\ldots,\alpha_k\rangle$ is ordered in R implies that $\mathtt{did}_{j-1}(\alpha_{j-1})$ is a necessary condition for $\mathtt{does}_j(\alpha_j)$ in the system R, and so $(R,r,\hat{t})\models\mathtt{did}_{j-1}(\alpha_{j-1})$. Since $\mathtt{did}_{j-1}(\alpha_{j-1})$ is a stable fact and $t\ge\hat{t}$, we obtain that $(R,r,t)\models\mathtt{did}_{j-1}(\alpha_{j-1})$. The claim follows. □

We say that a fact φ is ***stable in R*** if once true, φ remains true. Formally, if $(R,r,t)\models\varphi$ and $t'>t$ then $(R,r,t')\models\varphi$, for all $r\in R$ and $t,t'\ge 0$. Notice that while $\mathtt{does}_i(\alpha)$ is, in general, not a stable fact, $\mathtt{did}_i(\alpha)$ is always stable.

Definition 5.2. *We say that* ***agent i recalls ψ in R*** *if the fact* $K_i\psi$ *is stable in R.*

The notion of *perfect recall*, capturing the assumption that agents remember all events that they take part in, is popular in the analysis of games and multi-agent systems [11, 29]. While perfect recall is a nontrivial assumption often requiring significant storage costs, selective recall of single facts such as $\mathtt{does}_j(\alpha_j)$ is a much weaker assumption, that can be assumed of a system R essentially without loss of generality. By adding a single bit to Agent j's local state, whose value is 0 as long as j has not performed α_j and 1 once the action has been performed, we can obtain a system R' that is isomorphic to R, in which Agent j recalls $\mathtt{does}_j(\alpha_j)$.

Claim 2. *Assume that* α_j *is a conscious action for j in R, and that j recalls* $\mathtt{did}_j(\alpha_j)$ *in R. Then* $\mathtt{did}_j(\alpha_j)$ *is a j-local fact in R.*

Proof. Suppose that $(R,r,t)\models\mathtt{did}_j(\alpha_j)$. Then, by definition of $\mathtt{did}_j(\alpha_j)$, we have $(R,r,\hat{t})\models\mathtt{does}_j(\alpha_j)$ for some $\hat{t}\le t$. Choose an arbitrary $(r',t')\in\mathsf{Pts}(R)$ satisfying that $(r',t')\approx_j(r,\hat{t})$. Then $(R,r',t')\models\mathtt{does}_j(\alpha_j)$ since α_j is a conscious action for j in R. By definition of $\mathtt{did}_j(\alpha_j)$ it follows that $(R,r',t')\models\mathtt{did}_j(\alpha_j)$. Now, by definition of $\models$ for K_j we have that $(R,r,\hat{t})\models K_j\mathtt{did}_j(\alpha_j)$. By assumption, j recalls $\mathtt{did}_j(\alpha_j)$ in R, and so $K_j\mathtt{did}_j(\alpha_j)$ is stable in R. Thus, since $t\ge\hat{t}$, we obtain that $(R,r,t)\models K_j\mathtt{did}_j(\alpha_j)$, as claimed. □

We can now show:

Theorem 5.3 (Ordering and Nested Knowledge). *Assume that*

- *the actions* $\langle\alpha_1,\ldots,\alpha_k\rangle$ *are* ordered *in R,*
- *each agent* $j=1,\ldots,k$ *recalls* $\mathtt{did}_j(\alpha_j)$ *in R,*
- α_j *is a conscious action for j in R, for all* $j=1,\ldots,k$*, and*
- ψ *is a* stable *necessary condition for the first action* $\mathtt{does}_1(\alpha_1)$ *in R*

Then $K_jK_{j-1}\cdots K_1\psi$ *is a necessary condition for the* j^{th} *action* $\mathtt{does}_j(\alpha_j)$ *in R, for all* $j\le k$*.*

Proof. Assuming the conditions of the theorem, we will prove by induction on $j\le k$ that $\mathtt{did}_j(\alpha_j)$ validly implies $K_jK_{j-1}\cdots K_1\psi$ in R. Since $\mathtt{does}_j(\alpha_j)$ validly implies $\mathtt{did}_j(\alpha_j)$ by definition of $\mathtt{did}_j(\alpha_j)$, this will yield that $K_jK_{j-1}\cdots K_1\psi$ is a necessary condition for $\mathtt{does}_j(\alpha_j)$ in R, as claimed. We proceed with the inductive argument.

- Base case $j=1$: Assume that $(R,r,t)\models\mathtt{did}_1(\alpha_1)$. By Claim 2 we have that $(R,r,t)\models K_1\mathtt{did}_1(\alpha_1)$. Let $(r',t')\in\mathsf{Pts}(R)$ be an arbitrary point satisfying that $(r',t')\approx_1(r,t)$. Then $(R,r',t')\models\mathtt{did}_1(\alpha_1)$ by the knowledge property. Thus, $(R,r',\hat{t})\models\mathtt{does}_1(\alpha_1)$ holds for some $\hat{t}\le t'$, and because ψ is a necessary condition for $\mathtt{does}_1(\alpha_1)$ in R, we obtain that $(R,r,\hat{t})\models\psi$. Since ψ is stable and $t'\ge\hat{t}$, we have that $(R,r',t')\models\psi$. By choice of (r',t') we have that $(R,r,t)\models K_1\psi$, as claimed.

- Inductive step: Let $j > 1$ and assume that $K_{j-1} \cdots K_1\psi$ is a necessary condition in R for $\mathtt{did}_{j-1}(\alpha_{j-1})$. Moreover, let $(R, r, t) \models \mathtt{did}_j(\alpha_j)$. Since α_j is a conscious action for j, Claim 2 implies that $(R, r, t) \models K_j\mathtt{did}_j(\alpha_j)$. Choose an arbitrary $(r', t') \in \mathsf{Pts}(R)$ satisfying that $(r', t') \approx_j (r, t)$. By definition of K_j, it follows that $(R, r', t') \models \mathtt{did}_j(\alpha_j)$. By Claim 1, since the sequence $\langle\alpha_1, \ldots, \alpha_k\rangle$ is ordered in R and $j > 1$ we have that $(R, r', t') \models \mathtt{did}_{j-1}(\alpha_{j-1})$. We now apply the inductive hypothesis to obtain that $(R, r', t') \models K_{j-1} \cdots K_1\psi$. Finally, we obtain that $(R, r, t) \models K_jK_{j-1} \cdots K_1\psi$ by choice of (r', t') and the definition of '$\models$' for K_j. The claim now follows.

□

A slightly more restricted version of Theorem 5.3 was proved in [5]. Rather than consider an arbitrary necessary condition for α_1, they proved a version for the case in which the first action α_1 is triggered by an external input to agent 1. Technically, the proofs are quite similar.

Theorem 5.3 provides a necessary, but possibly not sufficient, condition for ordering actions in distributed systems. If agent j acts strictly later than when $K_jK_{j-1} \cdots K_1\psi$ first holds, then it may be inappropriate for agent $j + 1$ to act when it knows that the fact $K_jK_{j-1} \cdots K_1\psi$ holds (i.e., when $K_{j+1}K_j \cdots K_1\psi$ first holds). Nevertheless, Theorem 5.3 is often very useful because it can be used as a guide for efficiently, and sometimes even optimally, performing a sequence of ordered actions. Intuitively, suppose that we have a protocol whose goal is to perform $\langle\alpha_1, \ldots, \alpha_k\rangle$ in response to an externally generated trigger ψ (such as the '$\mathtt{go}$' message in Firing Squad). In particular, assume that ψ is a necessary condition for α_1. Keeping the communication aspects of this protocol fixed, an optimally fast solution would be for each agent $j \leq k$ to perform α_j when $K_jK_{j-1} \cdots K_1\psi$ first holds. Let R be the set of runs of such a protocol with $r \in R$, and let t_j and t_{j-1} be the earliest times at which $(R, r, t_j) \models K_jK_{j-1} \cdots K_1\psi$ and $(R, r, t_{j-1}) \models K_{j-1} \cdots K_1\psi$ hold in a run r, respectively. The knowledge property guarantees that $K_jK_{j-1} \cdots K_1\psi$ validly implies that $K_{j-1} \cdots K_1\psi$ in R, and so $t_j \geq t_{j-1}$. Since, by assumption, α_j is performed at time t_j and α_{j-1} at t_{j-1}, we have that agents perform actions in linear temporal order, as required by Definition 5.1. Clearly, none of the actions can be performed any earlier, as Theorem 5.3 shows. We conclude that in time-efficient protocols, the nested knowledge formula presented by the theorem can be both necessary and sufficient. In this sense, Theorem 5.3 suggests a recipe for obtaining time-efficient solutions for ordering actions.

Just as Theorem 4.3 implies that common knowledge is a necessary condition for simultaneous actions, we now have by Theorem 5.3 that nested knowledge is a necessary condition for performing actions in linear temporal order. And just as there is an established literature on when common knowledge is and is not attainable and on how it may arise, there are results concerning the communication structure that underlies attaining nested knowledge. Indeed, in a seminal paper [7], Chandy and Misra showed that in asynchronous systems R, if $(R, r, t) \models \neg\varphi$ and at a time $t' > t$ $(R, r, t') \models K_jK_{j-1} \cdots K_1\varphi$, then there must be a message chain in the run r between times t and t', passing through the agents 1,2,. . . ,j in this order (possibly involving additional agents as well). Given Theorem 5.3, this implies that the only way to coordinate actions in a linear temporal order in an asynchronous setting is by way of such message chains.[2]

More recently, Ben Zvi and Moses extended Chandy and Misra's work to systems in which communication is not asynchronous, but rather agents may have access to clocks and the transmission time for each of the channels is bounded [5]. They show that a communication structure called a *centipede* must be constructed in order to obtain nested knowledge of spontaneous facts such as the arrival of an external input. They prove a slightly more restricted instance of Theorem 5.3 (without using **K*o*P** directly), and use it to show that ordering actions in their setting requires the construction of the appropriate centipedes. Finally, Parikh and Krasucki analyze the ability to create *levels of knowledge* consisting of collections of nested knowledge formulas in [27]. Theorem 5.3 relates levels of knowledge to coordination.

6. DISCUSSION

This paper formulated the knowledge of preconditions principle and presented three theorems relating knowledge and coordinated action: the first is the **K*o*P** itself—necessary conditions for an action must be *known* to hold when the action is performed. Next, we showed that necessary conditions for simultaneous actions must be commonly known when the actions are taken. Finally, nested knowledge is a necessary condition for coordinating linearly ordered actions. The latter two are fairly direct consequences of the **K*o*P**. We discussed some of the uses of the latter two results in Sections 4 and 5. Indeed the **K*o*P** has many further implications.

In recent years, several works that make use of **K*o*P** have appeared, citing the unpublished [22]. For example, Castañeda, Gonczarowski and Moses used the **K*o*P** to analyze the consensus problem [6], in which agents need to agree on a binary value in a fault-prone system. They designed a protocol in two steps—applying the **K*o*P** once to derive a rule by which, roughly, agents decide on 0 when they know of an initial value of 0. Then, they applied the **K*o*P** again assuming this rule for decisions on 0, and obtained a rule involving nested knowledge (roughly, a statement of the form "knowing that nobody knows 0") for deciding on a value of 1. The result of their analysis was a very efficient solution to consensus that is optimal in a strong sense: It is the first unbeatable consensus protocol. The work of [6] complements an earlier work by Halpern, Moses and Waarts [15], in which a fixed point analysis of optimal consensus was obtained. It, too, is closely related to the **K*o*P**.

Gonczarowski and Moses used the **K*o*P** to analyze the epistemic requirements of more general forms of coordination [12]. Namely, they considered a setting in which k agents need to perform actions, and there are time bounds on the relative times at which the actions of any pair of agents is performed. The simple instance in which all bounds are 0 is precisely that of the simultaneous actions considered in Section 4. They show that such coordination requires vectorial fixed points of knowledge conditions, which are natu-

[2] Theorems 3.1 and 5.3 depend on conscious actions and therefore do not apply to asynchronous systems. Nevertheless, variants of these theorems can be presented that do apply to asynchronous systems and nondeterministic protocols. Details will appear in [22].

rally related to fixed points and equilibria. The papers [3, 4, 5, 12] together can all be viewed as making use of the **K*o*P** to provide insights into the interaction between time and communication for coordinating actions in a distributed and multi-agent system. Describing them is beyond the scope of the current paper.

The most significant aspect of the **K*o*P**, in our view, is the fact that it places a new emphasis on the epistemic aspects of problem solving in a multi-agent system. Simple necessary conditions induce epistemic conditions. Thus, in order to act correctly, one needs a mechanism ensuring that the agents obtain the necessary knowledge, and that they discover that they have this knowledge. Most problems and solutions are not posed or described in this fashion. We believe that the **K*o*P** encapsulates an important connection between knowledge, action and coordination that will find many applications in the future.

7. REFERENCES

[1] H. Attiya and J. Welch. *Distributed Computing: Fundamentals, Simulations and Advanced Topics.* John Wiley & Sons, 2004.

[2] R. J. Aumann. Agreeing to disagree. *Annals of Statistics*, 4(6):1236–1239, 1976.

[3] I. Ben-Zvi and Y. Moses. Agent-time epistemics and coordination. In *Proceedings of the 5th ICLA*, pages 97–108, 2013.

[4] I. Ben-Zvi and Y. Moses. The shape of distributed coordination. In *Proceedings of TARK XIV*, pages 29–38, 2013.

[5] I. Ben-Zvi and Y. Moses. Beyond lamport's *Happened-before*: On time bounds and the ordering of events in distributed systems. *Journal of the ACM*, 61(2):13, 2014.

[6] A. Castañeda, Y. A. Gonczarowski, and Y. Moses. Unbeatable consensus. In *Proceedings of DISC*, pages 91–106. Springer, 2014.

[7] K. M. Chandy and J. Misra. How processes learn. *Distributed Computing*, 1(1):40–52, 1986.

[8] D. Dolev, E. N. Hoch, and Y. Moses. An optimal self-stabilizing firing squad. *SIAM Journal on Computing*, 41(2):415–435, 2012.

[9] C. Dwork and Y. Moses. Knowledge and common knowledge in a Byzantine environment: crash failures. *Information and Computation*, 88(2):156–186, 1990.

[10] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. Common knowledge revisited. In *proceedings of TARK VI*, pages 283–298, 1996.

[11] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge.* MIT Press, 2003.

[12] Y. Gonczarowski and Y. Moses. Timely common knowledge: Characterising asymmetric distributed coordination via vectorial fixed points. In *Proceedings TARK XIV*, pages 79–93, 2013.

[13] J. Y. Halpern and Y. Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3):549–587, 1990. A preliminary version appeared in *Proceedings of the 3rd ACM PODC*, 1984.

[14] J. Y. Halpern and Y. Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54:319–379, 1992.

[15] J. Y. Halpern, Y. Moses, and O. Waarts. A characterization of eventual Byzantine agreement. In *Proceedings of the 9th ACM PODC*, pages 333–346, 1990.

[16] M. P. Herlihy, Y. Moses, and M. R. Tuttle. Transforming worst-case optimal solutions for simultaneous tasks into all-case optimal solutions. In *Proceedings of the 30th ACM PODC*, pages 231–238, 2011.

[17] G. Le Lann. Distributed systems-towards a formal approach. In *IFIP Congress*, volume 7, pages 155–160. Toronto, 1977.

[18] D. Lewis. *Convention, A Philosophical Study.* Harvard University Press, 1969.

[19] J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In D. Michie, editor, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, Edinburgh, 1969.

[20] T. Mizrahi and Y. Moses. Continuous consensus via common knowledge. *Distributed Computing*, 20(5):305–321, 2008.

[21] T. Mizrahi and Y. Moses. Continuous consensus with ambiguous failures. In *Distributed Computing and Networking*, pages 73–85. Springer, 2008.

[22] Y. Moses. *Knowledge and Coordinated Action*, to appear, 2016.

[23] Y. Moses and M. R. Tuttle. Programming simultaneous actions using common knowledge. *Algorithmica*, 3:121–169, 1988.

[24] G. Neiger. Consistent coordination and continual common knowledge. Manuscript. 1990.

[25] G. Neiger and R. Bazzi. Using knowledge to optimally achieve coordination in distributed systems. In *Proceedings of TARK IV* pages 43–59.

[26] G. Neiger and M. R. Tuttle. Common knowledge and consistent simultaneous coordination. *Distributed Computing*, 6(3):334–352, 1993.

[27] R. Parikh and P. Krasucki. Levels of knowledge in distributed computing. *Sādhanā*, 17(1):167–191, 1992.

[28] F. B. Schneider. Implementing fault-tolerant services using the state machine approach: A tutorial. *ACM Computing Surveys (CSUR)*, 22(4):299–319, 1990.

[29] R. Selten. Reexamination of the perfectness concept for equilibrium points in extensive games. *International journal of game theory*, 4(1):25–55, 1975.

Learning from unrealized versus realized prices

Abstract

Kathleen Ngangoué
DIW Berlin
Mohrenstr. 58
10117 Berlin
kngangoue@diw.de

Georg Weizsäcker
Humboldt Universität zu Berlin and DIW Berlin
Spandauer Str. 1
10178 Berlin
weizsaecker@hu-berlin.de

ABSTRACT

Market prices reflect information about an asset's fundamental value. However, it is still not clear the extent to which traders are able to utilize this information. In particular, the market design in itself could affect inferences. The psychology literature provides general evidence that the strategy used in decision-making is contingent on the task's complexity. The way markets differ in their complexity may influence traders' decisions. Some markets, for instance, require traders to condition their strategy on a future price, some do not. Although this framing variation in market design should be irrelevant to rational traders, trading optimally when the price is unknown is cumbersome: traders must, in that case, condition their demand strategy on a future price, whose implicit information needs to be anticipated. In an experimental market with diverse information, we investigate the extent to which traders update on the price, in situations where orders are submitted *before* or *after* the price has realized.

We compare investors' behavior in three market settings: a simultaneous limit order market, a simultaneous price list treatment and a sequential market. All treatments have in common that the price reflects private information of another market participant. In the limit order (LO) treatment, participants observe their private signal and state their maximum willingness to pay by placing a limit order. In the price list (PL) treatment the task is discretized: participants choose between buying and selling for a given list of possible price values. Both treatments, LO and PL, are considered as simultaneous because the market price realizes after participants submit their order. In the sequential (S) treatment, participants observe the market price *before* deciding to buy or to sell. Theoretically, decision outcomes should be identical at least in the treatments PL and S, where the strategy sets are isomorphic.

We find that subjects react strongly to their private information, but not to the price, in both simultaneous treatments LO and PL. In that sense, subjects appear to form naive beliefs. However, switching from a simultaneous system (LO, PL) to a sequential mechanism (S) improves inferences considerably. When the price is known at the time of bid submission, bids react to prices, to an extent that is roughly consistent with Bayesian updating. Surprisingly, in that case, subjects sometimes even overweight the information contained in the price. Hence, market designs affect how well agents process information and, therefore, how efficient prices become.

Continuous, lexicographic context dependence in a binary choice setting*

Patrick O'Callaghan[†]

ABSTRACT

When changes in environmental factors lead to a change in preference for one alternative over another, preference is context dependent. Continuous context dependence arises when appropriate perturbations of context preserve strict preference. For the binary choice setting, we characterise continuous context dependence via a function that is a utility at each context. We provide two examples of lexicographic context dependence that delineate the scope of the model. The first concerns reference dependence in a medical setting; the second considers a financial setting with unawareness and uncertainty.

Keywords
binary choice, context effects, lexicographic order, preferences, utility

1. INTRODUCTION

Cognitive psychology describes a context effect as the influence of environmental factors on our perceptions and responses. In the present paper, we formalise the effect of context on decision making via a context parameter. A decision maker's preferences are context dependent if strict preference for one alternative over another varies with context.

Context dependence of preference is supported by neuroscience which reveals a strong distinction between the intransigent nature of the structural networks of the brain on the one hand and the dynamic nature of its operational subnetworks on the other [15]. The presence of the latter is understood to highlight context dependence as an inherent property of cognitive brain activity.

Continuous dependence on context, or robustness of strict preference to perturbations in the context is a basic requirement. This property is argued to be crucial to brain activity by von Neumann in his unfinished work "The Computer and the Brain" [19]. This assumption is also present in many models of context dependence. Consider reference dependence in Kőszegy and Rabin [9]; menu dependence in Pattanaik and Xu [16]; status quo dependence in Masatlioglu and Ok [12] and Sagi [18]; memory dependence in Gilboa and Schmeidler [5]; belief dependence in Gilboa and Schmeidler [4].

In this paper we characterise continuous context dependence for the two alternative (binary choice) case. We show that, provided the set of contexts satisfies certain topological conditions, asymmetry of strict preference at each context and continuous context dependence are together sufficient for the existence of a function that is: firstly, a utility representation on alternatives at each context; and secondly, continuous on the set of contexts.

This representation theorem is a simplification of the main theorem of O'Callaghan [14]. In particular, since there are only two alternatives, we are able to provide a more basic and constructive proof that appeals to Urysohn's lemma instead of Michael's selection theorem. Similar to that paper, the topological conditions on the context space that we identify are just enough to yield the desired result. This allows us to study lexicographic context dependence.

Lexicographic context dependence arises when the set of contexts is lexicographically ordered independently of preferences on alternatives. The first example we consider in the present paper is a medical setting where the decision maker (henceforth Val) faces a choice between two medical procedures (alternatives) for her child. Similar to Kőszegy and Rabin [9], the contexts are reference points, though we assume that the reference points belong to a lexicographically ordered two dimensional set. The dominant dimension is the probability of survival and the other is a wealth related dummy variable (with values zero or one). Mathematically, this space is known as "double-arrow space" or "the split interval".

The second example is to a financial setting, where Val is a trader that must choose whether to buy or sell a certain financial derivative. Val may be both unaware and uncertain of the states of the world. Following Heifetz, Meier, and Schipper [7], we assume that unawareness is indexed by a complete lattice. The derivatives that Val trades are more profitable when the uncertainty is higher according to some measure of entropy. In contrast, we suppose that Val always prefers to be more aware. Indeed, because each unawareness level is associated with richer information, we suppose that Val lexicographically prefers higher awareness to higher uncertainty. When the cardinality of the set that indexes unawareness is uncountable and well-ordered, these assumptions yield a context space that is homeomorphic to the "extended long line". This connection allows us to identify preferences that satisfy continuous context dependence but have no continuous representation.

This latter example highlights the possibility that, when the set of contexts is too general, even though every repre-

* I thank Jeff Kline and John Quiggin for useful comments that helped to improve this paper.

[†] Department of Economics, University of Queensland, p.ocallaghan@uq.edu.au.

sentation of preferences is discontinuous on contexts, preferences are continuous. We deem this to be a modelling problem that is best avoided. For instance, if the set of awareness levels is assumed countable, this issue does not arise.

2. MODEL

Preferences at each context..

Let X denote a nonempty set. For each x in X, let $\prec_x$ denote a binary relation on the set $A = \{a, b\}$. The canonical interpretation of $a \prec_x b$ is that it formalises the statement "at context x, alternative b is strictly preferred to alternative a". Given the focus on binary choice a natural, alternative interpretation is that of stochastic choice. In that case, $a \prec_x b$ holds provided the empirical probability b is chosen is significantly different from $\frac{1}{2}$ in the forced choice setting of Fechner [3]. Our model accommodates either interpretation.

Context preferences..

The collection $\{\prec_x\}_{x\in X}$ of binary relations on A is the primitive object we refer to as *context preferences* or just *preferences*. Context preferences exclude any preference statements that Val may in fact be in a position to make regarding pairs of contexts, or indeed between one alternative-context pair and another. This information is intentionally ignored so that no assumption need be made concerning preferences over such objects.[1] Thus although our canonical interpretation is that of multiple epistemological viewpoints, one for each context, this is not a formal requirement.

Representing context dependence.

By a *representation* of context preferences, we mean a function of the form $U : A \times X \to \mathbb{R}$, where for any context $x \in X$, and any $i, j \in A$, j is strictly preferred to i at x if and only if $U(i, x) < U(j, x)$. That is, at each context x, there exists a utility function $U(\cdot, x)$ that represents preferences in the classical sense. When X is a singleton, preferences are context-free.

Continuous context dependence.

Context dependence is continuous at x if for each $i, j \in A$, j is strictly preferred to i at x, then the set of contexts X contains an open neighbourhood N of x such that for any other y in N, j is also strictly preferred to i at y. The fact that N is open means that the "direction of perturbation" is irrelevant, the important thing is that it is sufficiently small.

Characterising continuous context dependence.

$U : A \times X \to \mathbb{R}$ *characterises continuous context dependence* provided the function $U(i, \cdot)$ is continuous for each $i \in A$. Recall that this means that, for each i, the set $\{x : U(i, x) < r\}$ is open for every $r \in R$.

The set of contexts..

A set X is said to be a perfectly normal topological space if it satisfies the following three conditions.

0. For every x, the singleton set $\{x\}$ is closed.

1. Every pair of nonempty, closed and disjoint sets E and F, there exists a pair of open disjoint sets G and H such that $E \subseteq G$ and $F \subseteq H$.

2. Every closed set is can be written as a countable intersection of open sets.

If a set X satisfies 0, it is a T_0 topological space. The condition that ensures X is normal is 1, whilst 2 ensures X is perfect.

Example 1. *A familiar metrizable set is the usual set of nonnegative real numbers $\mathbb{R}_+$. Let the set of contexts X be the cartesian product of the set $\mathbb{N}$ of nonnegative integers with the half open interval of real numbers $[0, 1)$. Then $\mathbb{R}_+$ is topologically indistinguishable from $X = \mathbb{N} \times [0, 1)$ provided we endow the latter with the topology generated by the intervals of lexicographic order $<^*$: $k \times r <^* l \times t$ if and only if $k < l$ or $[k = l$ and $r < t]$, where $k, l \in \mathbb{N}$ and $r, t \in [0, 1)$. Since $\mathbb{R}_+$ is metrizable, so is X.*[2]

Recall that $\mathbb{N}$ is the smallest well-ordered infinite set. Its supremum is the first infinite ordinal number, and is often denoted by ω, and so we may write $\mathbb{N} \equiv [0, \omega)$. An extension of this is the well-ordered set $[0, \omega_1)$ of all countable ordinal numbers. ω_1 denotes the first uncountable ordinal. With the above lexicographic order $<^*$ and corresponding topology, the product $X' = [0, \omega_1) \times [0, 1)$ is perfectly normal, but not metrizable and is known as the "long line". As such, X' is a valid example of a context space for the purposes of the main theorem.

Extending yet further, the extended long line "glues" the point $\omega_1 \times 0$ to the upper end of the long line to obtain the set X''. (This is similar to considering the one-point compactification of $\mathbb{R}_+$ that yields the nonnegative extended real line). It is not hard to show the X'' is normal, but the following argument demonstrates that it is not perfect.

Let $\{G_n : n \in \mathbb{N}\}$ be an arbitrary collection of open neighbourhoods of $\omega_1 \times 0$ in X''. Then, by the definition of neighbourhood, each G_n contains an open $<^*$-interval of the form $(x^n, \omega_1 \times 0]$ for each n. For each $k < \omega_1$, there are uncountably many $l < \omega_1$ such that $k < l$. Let k_n satisfy $x^n < k_n \times 1 < \omega_1 \times 0$ for each n. Then the set $K = \{k_n : n \in \mathbb{N}\}$ is countable. By Munkres [13, Theorem 10.3], K has an upper bound $\bar{k} < \omega_1$. Then $\bigcap\{G_n : n \in \mathbb{N}\} \neq \{\omega_1 \times 0\}$ because $(\bar{k} \times 1, \omega_1 \times 0] \subset G_n$ for every n.

Note that in example 1, the order $<^*$ is over contexts X. The same is true of the examples in subsection 4.1 and 4.2. It is important to note, however, that $<^*$ serves only to generate interesting and familiar examples of topological spaces. In contrast, the family of context preference relations $\{\prec_x\}_{x\in X}$ is the kernel of our model. Indeed, the following are the axioms on $\{\prec_x\}_{x\in X}$ that we seek to characterise.

Axioms. *For a topological space X, we will assume:*

Asy. *for every $x \in X$ and every $i, j \in A$, if $i \prec_x j$, then not $j \prec_x i$; and*

CD. *for each $i, j \in A$, $\{x \in X : i \prec_x j\}$ is open in X .*

[1] In the language of measurement theory (see d'Aspremont and Gevers [2] for a recent survey) context preferences are low in the information hierarchy.

[2] We adopt the product notation $x = k \times r$ of Munkres [13] for any element $x \in X$.

For each x such that neither $a \prec_x b$ nor $b \prec_x a$, we write $a \sim_x b$. By construction therefore, $\sim_x$ is symmetric: $a \sim_x b$ if and only if $b \sim_x a$. Similarly, the weak preference relation $\precsim_x$ that equals $\prec_x \cup \sim_x$ is complete on A. (Thus, for each x, either $a \precsim_x b$ or $b \precsim_x a$.)

(Asy) is commonly referred to as asymmetry. An immediate consequence of (Asy) is that $\sim_x$ is reflexive: $i \sim_x i$ for $i \in A$. Since $|A| = 2$, $\sim_x$ and $\prec_x$ are clearly transitive. (Recall that $\sim_x$ is transitive provided that, for any i, j, k in A, $i \sim_x j \sim_x k$ implies $i \sim_x k$.) This ensures that, for each x, $\precsim_x$ is both complete and transitive.

Continuous (context) dependence (CD), which was motivated in the introduction and discussed further below, characterises the stability of strict preference for one alternative over another. Crucially, this is stability with respect to a perturbation of the context. Together with (Asy), (CD) ensures that the sets $\{x : a \prec_x b\}$ and $\{x : b \prec_x a\}$ are separated sets, where $\{x : a \sim_x b\}$ is the set that separates them. This means that they are not only disjoint, but neither contains a limit point of the other. The remaining consequences are summarised in theorem 2 which now follows.

3. MAIN RESULT

Theorem 2. *For $A = \{a, b\}$ and X perfectly normal, preferences satisfy* (Asy) *and* (CD) *if and only if there exists $U : A \times X \to \mathbb{R}$ such that for every $x \in X$,*

1. *$U(\cdot, x)$ is a utility representation of $\prec_x$; and*

2. *U characterises continuous context dependence.*

OF 2. As usual, the sufficiency of the axioms is the main step in the proof. We begin with this argument.

Sufficiency.

Suppose both the sets $\{x : a \prec_x b\}$ and $\{x : b \prec_x a\}$ are nonempty, the remaining cases follow by an identical argument. By (Asy), these sets are disjoint. By (CD), these sets are open and $F = \{x : b \precsim_x a\}$ is closed.

Since X is perfect and F is closed, there exists a countable collection $\{G_n\}_{n \in \mathbb{N}}$ of open sets satisfying $\bigcap_1^\infty G_n = F$. Note that $X - G_n \subseteq \{x : a \prec_x b\}$, so that F and $X - G_n$ are closed and disjoint for each n. Since X is normal, the Urysohn lemma applies. For each n, this guarantees the existence of a continuous function $f_n : X \to \mathbb{R}$ such that $f_n(x) = 0$ on F and $f_n(x) = 1$ on $X - G_n$, and $0 \leqslant f_n(x) \leqslant 1$ otherwise. Now let $f = \sum_1^\infty 2^{-n} f_n$. Since the uniform limit of continuous functions is continuous, f is continuous. Moreover, $f(x) > 0$ if and only if $a \prec_x b$. By the same argument, there exists another continuous nonnegative function $g : X \to \mathbb{R}$ such that $g^{-1}(0) = \{x : a \precsim_x b\}$.

Let $U(a, x) = 0$ for each $x \in X$ and let

$$U(b, x) = \begin{cases} f(x) & \text{if } a \prec_x b, \\ -g(x) & \text{if } b \prec_x a, \\ 0 & \text{otherwise.} \end{cases}$$

The resulting function $U : A \times X \to \mathbb{R}$ satisfies conditions 1 and 2 of the theorem.

Necessity.

Suppose that $U : A \times X \to \mathbb{R}$ satisfies condition 1 of the theorem and that $i \neq j \in A$ and, for some $x \in X$, $U(i, x) < U(j, x)$. Clearly, by asymmetry of $<$ on $\mathbb{R}$, it cannot be that $U(j, x) < U(i, x)$. Since x was arbitrarily chosen, 1 is sufficient for (Asy).

Suppose that U also satisfies condition 2 of the theorem. Since the difference of two continuous functions is continuous, $U(i, \cdot) - U(j, \cdot)$ is a continuous function. This ensures that the set $G = \{x : U(i, x) - U(j, x) < 0\}$ is open. By condition 1, $G = \{x : i \prec_x j\}$ and (CD) holds. □

Remark 3. *The proof of theorem 2 highlights the difficulties that arise when X fails to be perfectly normal. In the example of subsection 4.2, we provide a concrete example of preferences that fail to have a continuous representation even though they satisfy (CD). This observation is true of any context space that fails to be perfectly normal [14].*

4. APPLICATIONS

We now apply the model to an example of reference-dependent decision making in a medical setting followed by an example of unawareness in a financial setting.

4.1 Medical decision making

The main example of this subsection shows that the results apply to problems that cannot be modelled using either a context-free utility function or the previous results in the literature on jointly continuous utility representations. Extending it provides a concrete example where there is no continuous representation of preferences that satisfy continuous context dependence at the end of the subsection.

Contexts.

Consider a setting where Val is the mother of a child with a life threatening, but curable illness. Each context is a reference point in the sense of Kahneman and Tversky [8], but like Kőszegy and Rabin [9], the reference point is allowed to vary. Assume the set of reference points Val might face is $X = (0, 1) \times \{0, 1\}$. The first dimension $0 < p < 1$ denotes the (objective) probability that no medical complications arise from the procedure. We suppose that p would be provided by the child's physician on the basis of historical statistics and that, in this respect, there is no difference between the private and public sectors. The second dimension is a dummy variable that indicates whether Val is a home owner ($k = 1$) or not ($k = 0$).

Preferences.

Val is to choose whether or not she will go private (possibly taking out a loan to do so). The elements of A are therefore a ="go public" and b ="go private" respectively. The private sector provides a private room for her child and allows for arbitrarily long visiting hours. On the other hand b incurs a fixed cost of \$10,000.[3] There are no further (financial) costs relating to the complications that arise following the procedure. Wishing to be at her child's bedside in the case of complications, Val will always go private if $p < \frac{3}{4}$. If $\frac{99}{100} < p$, Val considers the risk of complications to be sufficiently low to warrant going public. Let $\underline{p} = \frac{3}{4}$ and $\bar{p} = \frac{99}{100}$. For the remaining contexts $\{x \in X : \underline{p} \leqslant x_1 \leqslant \bar{p}\}$, Val

[3] The setting we have in mind is the Australian healthcare system, where going private entails paying a portion of the doctors' fees for those with private health insurance. In that system, those with private health insurance may still choose to go public.

is indecisive between the two alternatives.[4] Thus, at each context $x \in X$, $\prec_x$ satisfies (Asy).

Lexicographic ordering of contexts.

As in example 1, we adopt the product notation $x = p \times k$ for elements of X. Since both dimensions of X are variables over which Val has no control, they are treated as context variables. Nonetheless, it may be reasonable to assume that Val has a strict preference ordering $<^*$ over X for contexts $x = p \times k$ with higher values of p regardless of the value of k. It may also reasonable to assume that, if $x = p \times k$ and $y = q \times l$ and $p = q$, then $x <^* y$ is better if $k < l$.

Continuous reference dependence.

Consider the following collection of (open) order intervals in the set of contexts X:

$$\{x : y <^* x\} \quad \text{and} \quad \{x : x <^* z\}$$

such that $y, z \in X$. The collection of open sets τ_X formed by taking unions of finite intersections of such intervals generates the lexicographic order topology on X. Note that, since $\{y : \underline{p} \times 0 <^* y <^* \underline{p} \times 1\}$ is empty, $\{y : \underline{p} \times 0 <^* y\} = \{y : \underline{p} \times 1 \leqslant^* y\}$. This ensures that the latter set is both open and closed in X. By the same argument, the the same is true of $\{y : y \leqslant^* \bar{p} \times 0\}$. Since the intersection of two open sets is open, $\{y : \underline{p} \times 1 \leqslant^* y \leqslant^* \bar{p} \times 0\}$ is both open and closed in X. Moreover, this set can also be rewritten as

$$\Big((\underline{p}, \bar{p}] \times \{0\} \Big) \cup \Big([\underline{p}, \bar{p}) \times \{1\} \Big).^5 \tag{1}$$

Lemma 4. *The set $\{x : a \sim_x b\}$ is closed in X.*

Proof. Since $\{y : a \sim_x b\} = \{y : \underline{p} \times 1 \leqslant^* y \leqslant^* \bar{p} \times 0\} \cup \{\underline{p} \times 0\} \cup \{\bar{p} \times 1\}$ and the union of three closed sets is closed, the proof is complete. □

Representation of preferences.

(X, τ_X) is a well known example of a perfectly normal topological space [6]. Thus, theorem 2 and lemma 4 guarantees the existence of a family of utility representations, one for each $\prec_x$ such that $x \in X$ that is continuous on X.

The space X is not metrizable, and so the results of Levin [11] and Caterino, Ceppitelli, and Maccarino [1] do not apply. See O'Callaghan [14] for a discussion of the connections with those results. As with many lexicographically ordered sets, $(<^*, X)$ cannot be represented by any context-free utility function with values in $\mathbb{R}$. This implies that Val's preferences cannot be extended to a context-free binary relation on $A \times X$ that contains $<^*$ and has a real-valued utility representation.

4.2 Unawareness frames

Consider a setting where Val is a financial derivatives trader who must choose between two alternatives in the face of both unawareness and uncertainty regarding the state of the world.

The context space.

In order to construct the set of contexts X, we consider a lattice L of standard state spaces as in Heifetz, Meier, and Schipper [7]. Each state space $S \in L$ represents a different level of awareness. We focus on the special case where the lattice order $\leqslant^L$ on L is well-ordered: that is, every $L' \subseteq L$ has a smallest element. Essentially, this restricts attention to the single agent unawareness problem, for then every pair of elements $S, S' \in L$ are comparable.

Heifetz, Meier, and Schipper [7] assume that L is a complete lattice. That is, every $L' \subseteq L$ has a greatest lower bound and a least upper bound according to $\leqslant^L$. As L is well-ordered, if it has a greatest element [10], then it is a complete lattice. Let ω_1 be the smallest uncountable ordinal number of example 1 and recall that $[0, \omega_1)$ is the set of all countable ordinals and $[0, \omega_1] = [0, \omega_1) \cup \{\omega_1\}$. Then $[0, \omega_1]$ is a complete lattice. Let 0 and ω_1 index the minimal and maximal elements of L respectively, and let $[0, \omega_1]$ and L be order isomorphic. This allows us to identify L and $[0, \omega_1]$, and where necessary drop reference to $S_k, S_l \in L$ and simply refer to $k, l \in L$.

Let L denote the first dimension of the context space. The second dimension represents different levels of uncertainty given the level of awareness. To keep the example to a minimum, we assume that each $S \in L$ is associated with a single σ-algebra Σ_S on S that represents Val's information. Moreover, we suppose that Val is only concerned a given entropy ranking $\leqslant^S$ of some subset D of the of set all of probability measures on Σ_S. In particular, for $p, q \in D$, $p \leqslant^S q$ means that the entropy of p is weakly lower than the entropy of q according to $\leqslant^S$. Recall that the entropy of a measure evaluates the disorder associated with a measure, for instance, uniform distributions have maximum entropy, and Dirac measures that assign full mass to any given point have minimum entropy.

It is standard to associate entropy orders with some interval in $\mathbb{R}$. We assume that $(D, \leqslant^S)$ is order isomorphic to the usual order $<$ on the half-open interval $[0, 1)$ in $\mathbb{R}$. This implies that the minimum entropy is attained for some probability measure in D, whereas the maximum entropy is not.

Remark 5. *Whilst the main reason for this assumption is to simplify the exposition, examples where the maximum entropy does not exist are provided by [17]. In particular, for probability measures with unbounded support, if the ratio of the third (skewness) and fourth (kurtosis) moments of a distribution is is high the maximum entropy does not exist. Alternatively, in a setting where the distributions have bounded support, and the uniform distribution has the maximum entropy, this assumption is equivalent to assuming Val always has some useful information, and this moves her away from the maximum entropy.*

Lexicographic ordering of contexts.

For certain trading strategies, uncertainty is a good thing. This is where the trader profits from higher volatility. We assume this is the case for Val. On the otherhand, because lower unawareness can imply trading with insiders, we assume higher awareness is always better for Val. So as to highlight the continuity issue that can arise in the simplest possible terms, we assume that $X = [0, \omega_1] \times [0, 1)$, and that Val lexicographically orders X as follows: if $x = k \times r$ and

[4] Being indecisive is observationally equivalent to being indifferent in this case.

[5] This property has lead $X \cup \{0, 1\}^2$ to be called "double-arrow" space.

$y = l \times t$, then $x \leqslant^* y$ if and only if $k <^L l$ or $[k = l$ and $r \leqslant^S t]$. As in the example of section 4.1, we consider the order topology on X associated with $<^*$.

Preferences given context.

Let a="sell" and b="buy". We assume that Val's preferences on $A = \{a, b\}$ are such that $a \sim_x b$ if and only if $x = \omega_1 \times 0$. For all $x <^* \omega_1 \times 0$, Val strictly prefers to buy at x. Finally, suppose that $b \prec_x a$ holds only if $\omega_1 \times 0 <^* x$. In other words, it is only when Val is fully aware and there is some uncertainty that she realises she would strictly prefer to be a seller. In the language of logic and Heifetz, Meier, and Schipper [7] in particular, there is some special proposition that belongs only to ω_1 that causes Val to prefer to sell.

Note that Val's preferences satisfy both (Asy) and (CD). The latter follows simply because the set $\{x : a \sim_x b\}$ contains only the point $\omega_1 \times 0$ and, by condition 0 of the definition of a perfectly normal set, the singleton sets are closed in the order topology generated by $<^*$. Nonetheless, theorem 2 does not apply because X is not perfectly normal.

No continuous characterisation.

Recall from example 1, that the extended long line is not perfectly normal because it contains a closed set that is not equal to the intersection of some countable collection of open sets. We have chosen Val's preferences so that $\{x : b \precsim_x a\} = \{\omega_1\} \times [0, 1)$. This leads us to the following proposition.

Proposition 6. *There is no representation of Val's preferences that characterises continuous context dependence.*

OF PROPOSITION 6. Note that $\{x : b \precsim_x a\} = \{\omega_1\} \times [0, 1)$. Following the construction in the proof of theorem 2, suppose there exists a continuous, nonnegative function $g : X \to \mathbb{R}$ such that $g^{-1}(0) = \{\omega_1\} \times [0, 1)$. Now let $G_n = \{x : g(x) < n^{-1}\}$ for each $n \in \mathbb{N}$. Then by construction, $\bigcap\{G_n : n \in \mathbb{N}\} = \{\omega_1\} \times [0, 1)$. This, however, contradicts the proof in the final paragraph of example 1 that X is not perfect. □

5. SUMMARY

We have studied a simple model of context dependence for the binary choice setting. For this case, the main theorem provides a representation of preferences that depend continuously on a perfectly normal context space. This allowed us to consider two examples of lexicographic context dependence. The first example was developed in the setting of a medical decision. There the context space satisfied the conditions for the representation theorem. The second example looked at unawareness and uncertainty in a financial setting. In this case, the context space was not perfectly normal and preferences were such that no continuous representation is available.

References

[1] Alessandro Caterino, Rita Ceppitelli, and Francesca Maccarino. "Continuous utility functions on submetrizable hemicompact k-spaces". In: *Applied General Topology* 10 (2 2009), pp. 187–195.

[2] C. d'Aspremont and L. Gevers. "Social welfare functionals and interpersonal comparability". In: *Handbook of social choice and welfare*. Ed. by K. Arrow, A. Sen, and K. Suzumura. Vol. 1. 2002, pp. 459–541.

[3] Gustav T. Fechner. *Elemente der Psychophysik (2 Volumes)*. 2nd ed. Breitkopf & Hartel, 1889.

[4] Itzhak Gilboa and David Schmeidler. "A derivation of expected utility maximization in the context of a game". In: *Games and Economic Behavior* 44.1 (2003), pp. 172 –182. ISSN: 0899-8256. DOI: `DOI : 10 . 1016 / S0899-8256(03)00015-0`.

[5] Itzhak Gilboa and David Schmeidler. "Act similarity in case-based decision theory". In: *Economic Theory* 9.1 (1997), pp. 47–61.

[6] Gary Gruenhage and Justin Tatch Moore. "Perfect compacta and basis problems in topology". In: *Open problems in topology II* (2011), p. 151.

[7] Aviad Heifetz, Martin Meier, and Burkhard C Schipper. "Interactive unawareness". In: *Journal of Economic Theory* 130.1 (2006), pp. 78–94.

[8] Daniel Kahneman and Amos Tversky. "Prospect theory: and analysis of decision under risk". In: *Econometrica* 47(2) (1979), pp. 263–291.

[9] Botond Kőszegy and Matthew Rabin. "A model of reference-dependent preferences". In: *The Quarterly Journal of Economics* 121(4) (2006), pp. 1133–1165.

[10] Joachim Lambek. *Lectures on Rings and Modules*. Chelsea Publishing Series. Chelsea Publishing Company, 1986. ISBN: 9780828422833.

[11] Vladimir L. Levin. "A Continuous Utility Theorem for Closed Preorders on a σ-Compact Metrizable Space". In: *Soviet Math. Doklady* 28 (1983), pp. 715–718.

[12] Yusufcan Masatlioglu and Efe A Ok. "Rational choice with status quo bias". In: *Journal of Economic Theory* 121.1 (2005), pp. 1–29.

[13] James R. Munkres. *Topology*. 2nd. Prentice Hall, 2000. ISBN: 9780131816299.

[14] Patrick O'Callaghan. "Minimal conditions for parametric continuity of a utility representation". 2015.

[15] H.-J. Park and K. Friston. "Structural and Functional Brain Networks: From Connections to Cognition". In: *Science* 342.1238411 (2013).

[16] Prasanta K. Pattanaik and Yongsheng Xu. "On Dominance and Context-Dependence in Decisions Involving Multiple Attributes". In: *Economics and Philosophy* 28 (Special Issue 02 2012), pp. 117–132.

[17] Michael Rockinger and Eric Jondeau. "Entropy densities with an application to autoregressive conditional skewness and kurtosis". In: *Journal of Econometrics* 106.1 (2002), pp. 119–142.

[18] Jacob S. Sagi. "Anchored preference relations". In: *Journal of Economic Theory* 130.1 (2006), pp. 283 –295. ISSN: 0022-0531. DOI: `http://dx.doi.org/10.1016/j.jet.2005.01.009`.

[19] John von Neumann. *The computer and the brain*. Yale University Press, 1976.

When Do Types Induce the Same Belief Hierarchy? [*]

Andrés Perea
EPICENTER and Dept. of Quantitative Economics
Maastricht University
a.perea@maastrichtuniversity.nl

ABSTRACT

Harsanyi (1967–1968) showed how infinite belief hierarchies can be encoded by means of type structures. Such encodings, however, are far from unique: Two different types – possibly from two different type structures – may generate exactly the same belief hierarchy. In this paper we present a finite recursive procedure, the *Type Partitioning Procedure*, which verifies whether two types, from two potentially different finite type structures, induce the same belief hierarchy or not. Important is that the procedure does not make explicit reference to belief hierarchies, but operates entirely within the language of type structures.

Keywords

Types, belief hierarchies

1. INTRODUCTION

Belief hierarchies play a fundamental role in the modern analysis of games. In games with *incomplete information* – where players face uncertainty about the opponents' utilities – it is important to model what a player believes about his opponents' utility functions, what he believes about the opponents' beliefs about their opponents' utility functions, and so on (Harsanyi (1962, 1967–1968), Böge and Eisele (1979), Mertens and Zamir (1985), Ely and Pęski (2006), Dekel, Fudenberg and Morris (2007), Weinstein and Yildiz (2007) and others). But even in games with *complete information* – where the players' actual utility functions are transparent to everyone – belief hierarchies naturally enter the analysis when we investigate the belief a player has about his opponents' *choices,* the belief he has about the opponents' beliefs about their opponents' choices, and so on. Such belief hierarchies are the basis for many concepts in epistemic game theory, most of which build upon the central notion of *common belief in rationality* (Brandenburger and Dekel (1987), Tan and Werlang (1988)). For an overview of these concepts, see Brandenburger (2007), Perea (2012) and Dekel and Siniscalchi (2013).

One important practical problem with belief hierarchies is that these are *infinite* objects, with *infinitely* many layers of beliefs. It is thus impossible to explicitly write down a belief hierarchy, layer by layer, as there are infinitely many of these. But then, the question naturally arises: Is there a way to represent belief hierarchies in a compact and convenient way?

Harsanyi (1967–1968) gave a positive and elegant answer to this question. He focused on a setting in which the belief hierarchies concern only the players' *utilities,* but his construction has later been extended to situations where players also hold beliefs about the opponents' *choices.* The construction that Harsanyi proposed was very simple: For every player we define a set of types, and for every type we define a utility function, together with a probabilistic belief about the *opponents' types.* From this very simple construction we can then derive, for every type, a first-order belief about the opponents' utility functions, a second-order belief about the opponents' first-order beliefs, and so on. That is, for every type we can derive a *full belief hierarchy* on the players' possible utility functions in the game. This construction by Harsanyi was a major step forward, as it allowed us to encode infinite belief hierarchies about utilities in a very compact and convenient fashion.

Harsanyi's original idea can easily be adapted to a framework where players also hold beliefs about other features besides the players' utilities. Assume that every player faces a basic space of uncertainty, which can include the parameters determining the players' utility functions, the set of opponents' choices, and possibly some other features as well. Now, consider for every player a set of types, and associate to every type a probabilistic belief about the basic space of uncertainty *and* the opponents' types. Then, similarly to Harsanyi's construction, we can derive for every type a full belief hierarchy, specifying a first-order belief about the basic space of uncertainty, a second-order belief about the opponents' first-order beliefs about *their* basic spaces of uncertainty, and so on. This construction, which we call a *type structure*, thus allows us to encode infinite belief hierarchies about any set of parameters in a very compact way.

Such encodings, however, are far from unique: Two different types – possibly from two different type structures – may encode one and the same belief hierarchy. In view of this "multiplicity problem" we ask the following natural question in this paper: When do two types, from two potentially different finite type structures, induce the same belief hierarchy?

Checking this directly, by explicitly comparing their induced first-order beliefs, second-order beliefs, and so on, may be quite cumbersome as one needs to check for infinitely many levels of belief. Instead, this paper presents a finite recursive procedure, the *Type Partitioning Procedure,* which

[*] This paper is a substantially revised version of an earlier paper with the same title. I would like to thank Pierpaolo Battigalli, Eddie Dekel, Amanda Friedenberg, Willemien Kets, Christian Nauerz, Miklós Pintér and Elias Tsakas for very useful comments on the earlier version.

tells us precisely when two such types induce the same belief hierarchy, and when they do not. The procedure works as follows. If we compare two types from two *different* type structures, we start by merging the two type structures into one large type structure. In every round, the procedure generates for every player a *partition* of the set of types in the large type structure, where the partition in the current round will always be a *refinement* of the partition from the previous round. Since we restrict to type structures with finitely many types, this procedure will always terminate within finitely many rounds. The equivalence classes in the final partitions will then be exactly those groups of types that generate the same belief hierarchy. That is, two types generate the same belief hierarchy exactly when they belong to the same equivalence class in the final partition. We actually show a bit more in Theorem 2: We prove that for every n, two types share the same n-th order belief precisely when they belong to the same equivalence class of the partition produced in round n of the procedure. In that sense, the *Type Partitioning Procedure* provides a convenient and automated way to verify whether two types share the same belief hierarchy, or the same beliefs up to a fixed order n. Important, moreover, is that the *Type Partitioning Procedure* does not make any explicit reference to belief hierarchies – it operates entirely within the "language" of type structures.

We use the *Type Partitioning Procedure* to establish an interesting property of belief hierarchies which – we believe – is new. Suppose we compare two types – possibly from two different type structures – and let N be the total number of types in these two type structures. Then, we prove in Corollary 1 that these two types induce the same belief hierarchy exactly when they induce the same N-th order belief. To prove this result we use the property that the *Type Partitioning Procedure* will always terminate within at most N rounds. In particular, the smaller the number of types in the type structure, the less belief levels we must check in order to conclude that two types share the same belief hierarchy.

The outline of this paper is as follows. In Section 2 we introduce type structures and belief hierarchies, and show how we can derive belief hierarchies from types. In Section 3 we describe the *Type Partitioning Procedure,* illustrate it by means of an example, and show how it characterizes those types that share the same belief hierarchy. In Section 4 we show how the procedure can be used to test whether two types from *different* type structures generate the same belief hierarchy or not. In Section 5 we use the *Type Partitioning Procedure* to test properties of type *structures* rather than individual types. For instance, we use the procedure to check whether a type structure contains redundant types or not, or whether one type structure is *contained* in another type structure, in the sense that for every type in the first structure there is a type in the second structure that generates the same belief hierarchy. Section 6, finally, contains the proofs.

2. BELIEF HIERARCHIES AND TYPES

In this section we show how belief hierarchies can be encoded by means of a type structure, and how every type within a type structure can be "decoded" by deriving a full belief hierarchy from it.

2.1 Encoding Belief Hierarchies by Type Structures

Consider a finite set of agents I. Assume that each agent i faces a *basic space of uncertainty* $\mathcal{X}_i = (X_i, \Sigma_i)$, where X_i is an arbitrary set and Σ_i a σ-algebra on X_i. That is, (X_i, Σ_i) is a measurable space. The combination $\mathcal{X} = (\mathcal{X}_i)_{i\in I}$ of basic uncertainty spaces is called a *multi-agent uncertainty space.*

If the agents are the players in a game, the basic space of uncertainty for player i could, for instance, be the set of opponents' choice combinations, or the set of parameters determining the utility functions of the players, or even a combination of the two. The first scenario is the standard framework for games with complete information, the second scenario is Harsanyi's (1967–1968) original setting for games with incomplete information, whereas the last scenario is investigated in Böge and Eisele (1979) and Mertens and Zamir (1985), among others. The sets X_i could also include external events that cannot be influenced by the agents, as is the case in Böge and Eisele (1979).

A *belief hierarchy* for player i specifies a probability measure on $\mathcal{X}_i$ – the *first-order* belief, a probability measure on $\mathcal{X}_i$ and the opponents' possible first-order beliefs – the *second-order* belief, and so on. Following Harsanyi's (1967–1968) approach, we will encode such infinite belief hierarchies by means of *type structures.* In this paper we focus on type structures with *finitely* many types, which of course imposes restrictions on the possible belief hierarchies we can encode. Indeed, there are belief hierarchies which can simply not be encoded by type structures with finitely many types.

Definition 1 (Type Structure). *Consider a multi-agent uncertainty space $\mathcal{X} = (X_i, \Sigma_i)_{i\in I}$. A* ***finite type structure*** *for $\mathcal{X}$ is a tuple $\mathcal{T} = (T_i, b_i)_{i\in I}$ where, for every player i,*

(a) T_i is the finite set of types for player i, and

(b) $b_i : T_i \to \Delta(X_i \times T_{-i}, \hat{\Sigma}_i)$ is a mapping that assigns to every type t_i a probabilistic belief $b_i(t_i) \in \Delta(X_i \times T_{-i}, \hat{\Sigma}_i)$ on his basic uncertainty space and the opponents' type combinations.

Here, $T_{-i} := \times_{j\neq i} T_j$. For any measurable space $(Y_i, \hat{\Sigma}_i)$, we denote by $\Delta(Y_i, \hat{\Sigma}_i)$ the set of probability measures on $(Y_i, \hat{\Sigma}_i)$. In part (b) of the definition, we assume $\hat{\Sigma}_i$ to be the product σ-algebra on $X_i \times T_{-i}$ induced by the σ-algebra Σ_i on X_i and the discrete σ-algebra on the finite set T_{-i}.

2.2 From Type Structures to Belief Hierarchies

In the previous subsection we have introduced a type structure as a way to encode belief hierarchies. We will now show how to "decode" a type within a type structure, by deriving the full belief hierarchy it induces.

Consider a finite type structure $\mathcal{T} = (T_i, b_i)_{i\in I}$ for Γ. Then, every type t_i within $\mathcal{T}$ induces an infinite belief hierarchy

$$h_i(t_i) = (h_i^1(t_i), h_i^2(t_i), ...),$$

where $h_i^1(t_i)$ is the induced first-order belief, $h_i^2(t_i)$ is the induced second-order belief, and so on. We will inductively define, for every n, the n-th order beliefs induced by types t_i in $\mathcal{T}$, building upon the $(n-1)$-th order beliefs that have been defined in the preceding step. We start by defining the first-order beliefs.

For every player i, and every type $t_i \in T_i$, define the first-order belief $h_i^1(t_i) \in \Delta(X_i, \Sigma_i)$ by

$$h_i^1(t_i)(E_i) := b_i(t_i)(E_i \times T_{-i}) \text{ for all } E_i \in \Sigma_i.$$

Now, suppose that $n \geq 2$, and assume that the $(n-1)$-th order beliefs $h_i^{n-1}(t_i)$ have been defined for all players i, and every type $t_i \in T_i$. Let

$$h_i^{n-1}(T_i) := \{h_i^{n-1}(t_i) \mid t_i \in T_i\}$$

be the finite set of $(n-1)$-th order beliefs for player i induced by types in T_i. For every $h_i^{n-1} \in h_i^{n-1}(T_i)$, let

$$T_i[h_i^{n-1}] := \{t_i \in T_i \mid h_i^{n-1}(t_i) = h_i^{n-1}\}$$

be the set of types in T_i that have the $(n-1)$-th order belief h_i^{n-1}.

Let $h_{-i}^{n-1}(T_{-i}) := \times_{j \neq i} h_j^{n-1}(T_j)$, and for a given $h_{-i}^{n-1} = (h_j^{n-1})_{j \neq i}$ in $h_{-i}^{n-1}(T_{-i})$ let $T_{-i}[h_{-i}^{n-1}] := \times_{j \neq i} T_j[h_j^{n-1}]$.

We define the n-th order beliefs $h_i^n(t_i)$ as follows. Let Σ_i^{n-1} be the product σ-algebra on $X_i \times h_{-i}^{n-1}(T_{-i})$ induced by the σ-algebra Σ_i on X_i and the discrete σ-algebra on the finite set $h_{-i}^{n-1}(T_{-i})$. For every type $t_i \in T_i$, let the n-th order belief $h_i^n(t_i) \in \Delta(X_i \times h_{-i}^{n-1}(T_{-i}), \Sigma_i^{n-1})$ be given by

$$h_i^n(t_i)(E_i \times \{h_{-i}^{n-1}\}) := b_i(t_i)(E_i \times T_{-i}[h_{-i}^{n-1}]) \qquad (1)$$

for every $E_i \in \Sigma_i$ and every $h_{-i}^{n-1} \in h_{-i}^{n-1}(T_{-i})$.

Finally, for every type $t_i \in T_i$, we denote by

$$h_i(t_i) := (h_i^n(t_i))_{n \in \mathbb{N}}$$

the *belief hierarchy* induced by t_i.

3. TYPE PARTITIONING PROCEDURE

Suppose we consider a finite type structure for some multi-agent uncertainty space $\mathcal{X}$. In this section we will present a recursive procedure – the *Type Partitioning Procedure* – that tells us, within finitely many steps, which types from this type structure induce the same belief hierarchy and which do not. Important is that this procedure does not make explicit reference to belief hierarchies, but operates entirely within the "language" of type structures.

3.1 Definition of the Procedure

To formally define this procedure, we need the following terminology. A *finite partition* of a set A is a finite collection $\mathcal{P} = \{P_1, ..., P_K\}$ of nonempty subsets $P_k \subseteq A$ such that $\cup_{k=1}^K P_k = A$ and $P_k \cap P_m = \emptyset$ whenever $k \neq m$. We refer to the sets P_k as *equivalence classes.* For an element $a \in A$, we denote by $\mathcal{P}(a)$ the equivalence class P_k to which a belongs. The *trivial partition* of A is the partition $\mathcal{P} = \{A\}$ containing only one set – the full set A. For two partitions $\mathcal{P}^1$ and $\mathcal{P}^2$ on A, we say that $\mathcal{P}^1$ is a *refinement* of $\mathcal{P}^2$ if for every set $P^1 \in \mathcal{P}^1$ there is a set $P^2 \in \mathcal{P}^2$ such that $P^1 \subseteq P^2$. We say that $\mathcal{P}^1$ is a *strict refinement* of $\mathcal{P}^2$ if $\mathcal{P}^1$ is a *refinement* of $\mathcal{P}^2$ and $\mathcal{P}^1 \neq \mathcal{P}^2$.

In the procedure we recursively partition the set of types of an agent into equivalence classes – starting from the trivial partition, and refining the previous partition with every step – until these partitions cannot be refined any further. We show that the equivalence classes produced in round n contain exactly the types that induce the same belief hierarchy up to order n. In particular, the equivalence classes produced at the end contain precisely those types that induce the same (infinite) belief hierarchy.

Type structure $\mathcal{T} = (T_1, T_2, b_1, b_2)$
$T_1 = \{t_1, t_1', t_1''\}, \qquad T_2 = \{t_2, t_2', t_2''\}$
$b_1(t_1) = \frac{1}{2}(c, t_2) + \frac{1}{2}(d, t_2')$
$b_1(t_1') = \frac{1}{6}(c, t_2) + \frac{1}{3}(c, t_2'') + \frac{1}{2}(d, t_2')$
$b_1(t_1'') = \frac{1}{2}(c, t_2') + \frac{1}{2}(d, t_2'')$
$b_2(t_2) = \frac{1}{4}(e, t_1) + \frac{1}{2}(e, t_1') + \frac{1}{4}(f, t_1'')$
$b_2(t_2') = \frac{1}{8}(e, t_1) + \frac{1}{8}(e, t_1') + \frac{3}{4}(f, t_1'')$
$b_2(t_2'') = \frac{3}{8}(e, t_1) + \frac{3}{8}(e, t_1') + \frac{1}{4}(f, t_1'')$

Table 1: The type structure from Example 1

Procedure 1 (Type Partitioning Procedure). *We consider a multi-agent uncertainty space $\mathcal{X} = (X_i, \Sigma_i)_{i \in I}$, and a finite type structure $\mathcal{T} = (T_i, b_i)_{i \in I}$ for $\mathcal{X}$.*

Initial step. *For every agent i, let $\mathcal{P}_i^0$ be the trivial partition of his set of types T_i.*

Inductive step. *Suppose that $n \geq 1$, and that the partitions $\mathcal{P}_i^{n-1}$ have been defined for every agent i. Then, for every agent i, and every $t_i \in T_i$,*

$$\mathcal{P}_i^n(t_i) = \{t_i' \in T_i \mid b_i(t_i')(E_i \times P_{-i}^{n-1}) = b_i(t_i)(E_i \times P_{-i}^{n-1}) \qquad (2)$$
$$\text{for all } E_i \in \Sigma_i, \text{ and all } P_{-i}^{n-1} \in \mathcal{P}_{-i}^{n-1}\}.$$

The procedure ***terminates at round*** *n whenever $\mathcal{P}_i^n = \mathcal{P}_i^{n-1}$ for every agent i.*

In this procedure, $\mathcal{P}_{-i}^{n-1}$ is the partition of the set T_{-i} induced by the partitions $\mathcal{P}_j^{n-1}$ on T_j. More precisely, if $t_{-i} = (t_j)_{j \neq i}$ is in T_{-i}, then

$$\mathcal{P}_{-i}^{n-1}(t_{-i}) := \times_{j \neq i} \mathcal{P}_j^{n-1}(t_j),$$

which is a subset of T_{-i}.

We will now illustrate the *Type Partitioning Procedure* by means of an example.

Example 1. Consider a multi-agent uncertainty space $\mathcal{X} = (X_i, \Sigma_i)_{i \in I}$ where $I = \{1, 2\}$, $X_1 = \{c, d\}, X_2 = \{e, f\}$, and Σ_1, Σ_2 are the discrete σ-algebras on X_1 and X_2, respectively. Consider the type structure $\mathcal{T} = (T_1, T_2, b_1, b_2)$ in Table 1. Here, $b_1(t_1) = \frac{1}{2}(c, t_2) + \frac{1}{2}(d, t_2')$ means that type t_1 assigns probability $\frac{1}{2}$ to the pair $(c, t_2) \in X_1 \times T_2$, and probability $\frac{1}{2}$ to the pair $(d, t_2') \in X_1 \times T_2$. Similarly for the other types in the table. We will now run the *Type Partitioning Procedure.*

Initial Step. Let $\mathcal{P}_1^0$ be the trivial partition of the set of types T_1, and let $\mathcal{P}_2^0$ be the trivial partition of the set of types T_2. That is,

$$\mathcal{P}_1^0 = \{\{t_1, t_1', t_1''\}\} \text{ and } \mathcal{P}_2^0 = \{\{t_2, t_2', t_2''\}\}.$$

Round 1. By equation (2),

$$\begin{aligned}\mathcal{P}_1^1(t_1) = \{\tau_1 \in T_1 \mid \\ b_1(\tau_1)(\{c\} \times T_2) = b_1(t_1)(\{c\} \times T_2) = \tfrac{1}{2}, \\ b_1(\tau_1)(\{d\} \times T_2) = b_1(t_1)(\{d\} \times T_2) = \tfrac{1}{2}\} \\ = \{t_1, t_1', t_1''\},\end{aligned}$$

which implies that

$$\mathcal{P}_1^1 = \mathcal{P}_1^0 = \{\{t_1, t_1', t_1''\}\}.$$

At the same time,

$$\begin{aligned}\mathcal{P}_2^1(t_2) = \{\tau_2 \in T_2 \mid \\ b_2(\tau_2)(\{e\} \times T_1) = b_2(t_2)(\{e\} \times T_1) = \tfrac{3}{4}, \\ b_2(\tau_2)(\{f\} \times T_1) = b_2(t_2)(\{f\} \times T_1) = \tfrac{1}{4}\} \\ = \{t_2, t_2''\}\end{aligned}$$

which implies that $\mathcal{P}_2^1(t_2') = \{t_2'\}$, and hence

$$\mathcal{P}_2^1 = \{\{t_2, t_2''\}, \{t_2'\}.$$

Round 2. By equation (2),

$$\begin{aligned}\mathcal{P}_1^2(t_1) = \{\tau_1 \in T_1 \mid \\ b_1(\tau_1)(\{c\} \times \{t_2, t_2''\}) = b_1(t_1)(\{c\} \times \{t_2, t_2''\}) = \tfrac{1}{2}, \\ b_1(\tau_1)(\{c\} \times \{t_2'\}) = b_1(t_1)(\{c\} \times \{t_2'\}) = 0, \\ b_1(\tau_1)(\{d\} \times \{t_2, t_2''\}) = b_1(t_1)(\{d\} \times \{t_2, t_2''\}) = 0, \\ b_1(\tau_1)(\{d\} \times \{t_2'\}) = b_1(t_1)(\{d\} \times \{t_2'\}) = \tfrac{1}{2}\} \\ = \{t_1, t_1'\},\end{aligned}$$

which implies that $\mathcal{P}_1^2(t_1'') = \{t_1''\}$, and hence

$$\mathcal{P}_1^2 = \{\{t_1, t_1'\}, \{t_1''\}\}.$$

Since $\mathcal{P}_1^1 = \mathcal{P}_1^0$, we may immediately conclude that

$$\mathcal{P}_2^2 = \mathcal{P}_2^1 = \{\{t_2, t_2''\}, \{t_2'\}\}.$$

Round 3. As $\mathcal{P}_2^2 = \mathcal{P}_2^1$, we may immediately conclude that

$$\mathcal{P}_1^3 = \mathcal{P}_1^2 = \{\{t_1, t_1'\}, \{t_1''\}\}.$$

By equation (2),

$$\begin{aligned}\mathcal{P}_2^3(t_2) = \{\tau_2 \in T_2 \mid \\ b_2(\tau_2)(\{e\} \times \{t_1, t_1'\}) = b_2(t_2)(\{e\} \times \{t_1, t_1'\}) = \tfrac{3}{4}, \\ b_2(\tau_2)(\{e\} \times \{t_1''\}) = b_2(t_2)(\{e\} \times \{t_1''\}) = 0, \\ b_2(\tau_2)(\{f\} \times \{t_1, t_1'\}) = b_2(t_2)(\{f\} \times \{t_1, t_1'\}) = 0, \\ b_2(\tau_2)(\{f\} \times \{t_1''\}) = b_2(t_2)(\{f\} \times \{t_1''\}) = \tfrac{1}{4}\} \\ = \{t_2, t_2''\},\end{aligned}$$

which implies that $\mathcal{P}_2^3(t_2') = \{t_2'\}$, and hence

$$\mathcal{P}_2^3 = \{\{t_2, t_2''\}, \{t_2'\}\} = \mathcal{P}_2^2.$$

As $\mathcal{P}_1^3 = \mathcal{P}_1^2$ and $\mathcal{P}_2^3 = \mathcal{P}_2^2$, the procedure terminates at round 3. The final partitions of the types are thus given by

$$\mathcal{P}_1^\infty = \{\{t_1, t_1'\}, \{t_1''\}\} \text{ and } \mathcal{P}_2^\infty = \{\{t_2, t_2''\}, \{t_2'\}\}.$$

The reader may check that all types within the same equivalence class indeed induce the same belief hierarchy. That is, t_1 induces the same belief hierarchy as t_1', and t_2 induces the same belief hierarchy as t_2''. Moreover, t_1 and t_1'' induce different belief hierachies, and so do t_2 and t_2'. □

3.2 Characterization Result

We will now show that the *Type Partitioning Procedure* identifies precisely those types that share the same belief hierarchy. As a preparatory result, we will first highlight two important properties of the procedure. The first property states that the procedure is *monotonic* in the sense that the partitions generated at a particular round will always be *refinements* of the partitions generated in the round before. The second property states that the number of rounds that is needed for the procedure to terminate can never be larger than the total number of types we consider.

Theorem 1 (Properties of Procedure). *Consider a multi-agent uncertainty space $\mathcal{X} = (X_i, \Sigma_i)_{i\in I}$, and a finite type structure $\mathcal{T} = (T_i, b_i)_{i\in I}$ for $\mathcal{X}$. For every agent i and every round $n \geq 0$, let $\mathcal{P}_i^n$ be the partition of T_i generated in round n of the Type Partitioning Procedure. Let N be the total number of types in $\cup_{i\in I} T_i$. Then,*

(a) the partition $\mathcal{P}_i^n$ will always be a refinement of $\mathcal{P}_i^{n-1}$, for all agents i and all $n \geq 1$;

(b) the procedure will terminate after at most N rounds.

With this result at hand we can now prove the main theorem in this paper, which states that the *Type Partitioning Procedure* characterizes precisely those groups of types that induce the same belief hierarchy. We actually prove a little more: we show that the partitions generated in round n of the procedure characterize exactly those types that yield the same n-th order belief.

Theorem 2 (Characterization Result). *Consider a multi-agent uncertainty space $\mathcal{X} = (X_i, \Sigma_i)_{i\in I}$, and a finite type structure $\mathcal{T} = (T_i, b_i)_{i\in I}$ for $\mathcal{X}$. For every agent i and every round $n \geq 0$, let $\mathcal{P}_i^n$ be the partition of T_i generated in round n of the Type Partitioning Procedure. Let $\mathcal{P}_i^\infty$ be the final partition generated by the procedure. Then, for every agent i, every $n \geq 1$, and every two types $t_i, t_i' \in T_i$, we have that*

(a) $h_i^n(t_i) = h_i^n(t_i')$, if and only if, $t_i' \in \mathcal{P}_i^n(t_i)$;

(b) $h_i(t_i) = h_i(t_i')$, if and only if, $t_i' \in \mathcal{P}_i^\infty(t_i)$.

By combining Theorems 1 and 2 we can derive some interesting facts about finite type structures and belief hierarchies, which we state in the following corollary.

Corollary 1 (Properties of Belief Hierarchies). *Consider a multi-agent uncertainty space $\mathcal{X} = (X_i, \Sigma_i)_{i\in I}$, and a finite type structure $\mathcal{T} = (T_i, b_i)_{i\in I}$ for $\mathcal{X}$. Let N be the total number of types in $\cup_{i\in I} T_i$, and let $t_i, t_i' \in T_i$. Then,*

(a) for every $n \geq 2$, $h_i^{n-1}(t_i) = h_i^{n-1}(t_i')$ whenever $h_i^n(t_i) = h_i^n(t_i')$;

(b) $h_i(t_i) = h_i(t_i')$, if and only if, $h_i^N(t_i) = h_i^N(t_i')$.

Property (a) thus states that two types agreeing on the n-th order belief will also agree on all lower order beliefs. That is, the n-th order belief completely determines the first-order belief, the second-order belief, until the $(n-1)$-th order belief. Property (b) says that in order to check whether two types share the same infinite belief hierarchy or not we only have to compare the N-th order beliefs, where N is the total number of types in the type structure. To the best of our knowledge, this result is new in the literature.

The proof of this corollary is actually very easy. To show property (a) consider two types t_i, t_i' with $h_i^n(t_i) = h_i^n(t_i')$. Then, by Theorem 2, $t_i' \in \mathcal{P}_i^n(t_i)$. Since, by Theorem 1, $\mathcal{P}_i^n$ is a refinement of $\mathcal{P}_i^{n-1}$, it follows that $t_i' \in \mathcal{P}_i^{n-1}(t_i)$ and hence, by Theorem 2, $h_i^{n-1}(t_i) = h_i^{n-1}(t_i')$.

To show property (b), take two types t_i, t_i' with $h_i^N(t_i) = h_i^N(t_i')$. Then, by Theorem 2, $t_i' \in \mathcal{P}_i^N(t_i)$. By Theorem 1 we know that the procedure terminates after at most N rounds, and hence $\mathcal{P}_i^N = \mathcal{P}_i^\infty$. By Theorem 2 we conclude that $h_i(t_i) = h_i(t_i')$.

Some readers may ask why property (a) requires a proof, as in most other papers in the literature the n-th order belief induced by a type explicitly contains the $(n-1)$-th order belief as a component, and hence property (a) holds trivially. See, for instance, Heifetz and Samet (1998) and Friedenberg and Meier (2011). However, this is not the case in our setting: Our definition of $h_i^n(t_i)$ does not explicitly carry $h_i^{n-1}(t_i)$ as a component, and it is therefore not obvious that the n-th order belief fully determines the $(n-1)$-th order belief. This is a result, which requires a proof in our setting.

4. COMPARING TYPES FROM DIFFERENT TYPE STRUCTURES

We have seen that the *Type Partitioning Procedure* tells us exactly which types within a given type structure $\mathcal{T}$ induce the same belief hierarchy, and which do not. But what if we want to compare types from *different* type structures? We will see that the procedure will work for such settings as well.

Let us consider two different finite type structures, $\mathcal{T}^1 = (T_i^1, b_i^1)_{i \in I}$ and $\mathcal{T}^2 = (T_i^2, b_i^2)_{i \in I}$, for the same multi-agent uncertainty space $\mathcal{X} = (X_i, \Sigma_i)_{i \in I}$. For a given agent i, take a type $t_i^1 \in T_i^1$ and a type $t_i^2 \in T_i^2$. How can we check whether t_i^1 and t_i^2 induce the same belief hierarchy?

What we can do is to first merge the two type structures into one large type structure, and to subsequently run the *Type Partitioning Procedure* for the large type structure. More precisely, let $\mathcal{T} = (T_i, b_i)_{i \in I}$ be the "large" type structure, where $T_i := T_i^1 \cup T_i^2$ for all agents i, and

$$b_i(t_i) := \left\{ \begin{array}{ll} b_i^1(t_i), & \text{if } t_i \in T_i^1 \\ b_i^2(t_i), & \text{if } t_i \in T_i^2 \end{array} \right.$$

for all types $t_i \in T_i$. Hence, $\mathcal{T}$ is a "block" type structure in which types in T_i^1 only refer to opponents' types in T_{-i}^1, and types in T_i^2 only refer to opponents' types in T_{-i}^2. But it is still a well-defined type structure, and hence we can run the *Type Partitioning Procedure* for the "block" type structure $\mathcal{T}$, yielding partitions $\mathcal{P}_i^\infty$ of the sets $T_i = T_i^1 \cup T_i^2$ for every agent i. If $t_i^1 \in T_i^1$ and $t_i^2 \in T_i^2$ turn out to be in the same equivalence class of $\mathcal{P}_i^\infty$, then t_i^1 and t_i^2 induce the same belief hierarchy. Otherwise not. In this way, the *Type Partitioning Procedure* can also be used to test whether two types from different type structures induce the same belief hierarchy or not.

Type structure $\mathcal{T}^1 = (T_1^1, T_2^1, b_1^1, b_2^1)$

$T_1^1 = \{t_1, t_1', t_1'', t_1'''\}, \qquad T_2^1 = \{t_2, t_2', t_2''\}$

$$\begin{aligned}
b_1^1(t_1) &= \tfrac{1}{2}(c, t_2) + \tfrac{1}{2}(d, t_2') \\
b_1^1(t_1') &= \tfrac{1}{6}(c, t_2) + \tfrac{1}{3}(c, t_2'') + \tfrac{1}{2}(d, t_2') \\
b_1^1(t_1'') &= \tfrac{1}{2}(c, t_2') + \tfrac{1}{2}(d, t_2'') \\
b_1^1(t_1''') &= \tfrac{1}{3}(c, t_2) + \tfrac{2}{3}(d, t_2'')
\end{aligned}$$

$$\begin{aligned}
b_2^1(t_2) &= \tfrac{1}{4}(e, t_1) + \tfrac{1}{2}(e, t_1') + \tfrac{1}{4}(f, t_1'') \\
b_2^1(t_2') &= \tfrac{1}{8}(e, t_1) + \tfrac{1}{8}(e, t_1') + \tfrac{3}{4}(f, t_1'') \\
b_2^1(t_2'') &= \tfrac{3}{8}(e, t_1) + \tfrac{3}{8}(e, t_1') + \tfrac{1}{4}(f, t_1'')
\end{aligned}$$

Type structure $\mathcal{T}^2 = (T_1^2, T_2^2, b_1^2, b_2^2)$

$T_1^2 = \{r_1, r_1', r_1''\}, \qquad T_2^2 = \{r_2, r_2', r_2''\}$

$$\begin{aligned}
b_1^2(r_1) &= \tfrac{1}{4}(c, r_2) + \tfrac{1}{4}(c, r_2'') + \tfrac{1}{2}(d, r_2') \\
b_1^2(r_1') &= \tfrac{1}{2}(c, r_2') + \tfrac{1}{8}(d, r_2) + \tfrac{3}{8}(d, r_2'') \\
b_1^2(r_1'') &= \tfrac{1}{2}(c, r_2') + \tfrac{3}{8}(d, r_2) + \tfrac{1}{8}(d, r_2'')
\end{aligned}$$

$$\begin{aligned}
b_2^2(r_2) &= \tfrac{1}{4}(e, r_1') + \tfrac{3}{4}(f, r_1) \\
b_2^2(r_2') &= \tfrac{3}{4}(e, r_1') + \tfrac{1}{4}(f, r_1) \\
b_2^2(r_2'') &= \tfrac{1}{8}(e, r_1') + \tfrac{1}{8}(e, r_1'') + \tfrac{3}{4}(f, r_1)
\end{aligned}$$

Table 2: The type structures from Example 2

n	$\mathcal{P}_1^n$	$\mathcal{P}_2^n$
0	$\{\{t_1, t_1', t_1'', t_1''', r_1, r_1', r_1''\}\}$	$\{\{t_2, t_2', t_2'', r_2, r_2', r_2''\}\}$
1	$\{\{t_1, t_1', t_1'', r_1, r_1', r_1''\}, \{t_1'''\}\}$	$\{\{t_2, t_2'', r_2'\}, \{t_2', r_2, r_2''\}\}$
2	$\{\{t_1, t_1', r_1', r_1''\}, \{t_1'', r_1\}, \{t_1'''\}\}$	$\{\{t_2, t_2'', r_2'\}, \{t_2', r_2, r_2''\}\}$
3	$\{\{t_1, t_1', r_1', r_1''\}, \{t_1'', r_1\}, \{t_1'''\}\}$	$\{\{t_2, t_2'', r_2'\}, \{t_2', r_2, r_2''\}\}$

Table 3: The *Type Partitioning Procedure* in Example 2

Example 2. To see how this works, let us consider an example with two agents, $I = \{1, 2\}$, where the basic spaces of uncertainty are again given by $X_1 = \{c, d\}$ and $X_2 = \{e, f\}$ – as in Example 1 – together the the discrete σ-algebras on these sets. Consider the two type structures $\mathcal{T}^1 = (T_i^1, b_i^1)_{i \in I}$ and $\mathcal{T}^2 = (T_i^2, b_i^2)_{i \in I}$ on $\mathcal{X}$ as given in Table 2. Note that type structure $\mathcal{T}^1$ is almost identical to the type structure in Example 1, except for the fact that we have added an extra type t_1''' for agent 1.

We want to test whether the types $t_1 \in T_1^1$ and $r_1 \in T_1^2$, which belong to different type structures, induce the same belief hierrachy or not. As a first step we merge the two type structures $\mathcal{T}^1$ and $\mathcal{T}^2$ into one large block type structure, as described above. If we subsequently run the *Type Partitioning Procedure* for the large type structure, then the reader may verify that the partitions in every round are given by Table 3. Here, the procedure terminates at round 3. As t_1 and r_1 are not in the same equivalence class, we conclude that t_1 and r_1 do not induce the same belief hierarchy. In fact, the final partitions tell us that for agent 1, the types t_1, t_1', r_1' and r_1'' all induce the same belief hierarchy, that types t_1'' and r_1 induce the same belief hierarchy, and that for type $t_1''' \in T_1^1$ there is no type in T_1^2 that induces the same belief hierarchy. For agent 2, the types t_2, t_2'' and r_2' all induce the same belief hierarchy, and so do the types t_2', r_2 and r_2''. □

5. TESTING FOR PROPERTIES OF TYPE STRUCTURES

The *Type Partitioning Procedure* can be used to answer several different questions – local and global. First, as we already discussed, we can use it to test whether two types – possibly from different type structures – induce the same belief hierarchy or not. This is a local test.

But we can also use it to test global properties of type structures. For instance, we can use the procedure to test whether a given type structure contains *redundant* types or not – where "redundant" means that two different types induce the same belief hierarchy. For this redundancy test we can run the *Type Partitioning Procedure* and see whether the final partitions contain equivalence classes with at least two types. If this is the case then we conclude that the type structure contains redundancies. If, on the other hand, all equivalence classes contain only one type, then there are no redundancies in the type structure.

In case the type structure contains redundant types, the *Type Partitioning Procedure* will also tell us how to "remove" these redundancies without changing the induced collection of belief hierarchies. What we can do in this case is to replace every equivalence class in the final partition by a single type, and to change the belief of every type accordingly. Then we will obtain a smaller, non-redundant type structure that induces exactly the same collection of belief hierarchies.

As an illustration, consider the type structure from Table 1. We have seen in Example 1 that the *Type Partitioning Procedure* generates the final partitions

$$\mathcal{P}_1^\infty = \{\{t_1, t_1'\}, \{t_1''\}\} \text{ and } \mathcal{P}_2^\infty = \{\{t_2, t_2''\}, \{t_2'\}\}.$$

If we replace the equivalence class $\{t_1, t_1'\}$ by the single type r_1, and replace the equivalence class $\{t_2, t_2''\}$ by the single type r_2, then we obtain the smaller type structure $\hat{\mathcal{T}} = (\hat{T}_i, \hat{b}_i)_{i \in I}$ where

$$\hat{T}_1 = \{r_1, t_1''\},\ \hat{T}_2 = \{r_2, t_2'\}$$

and

$$\begin{aligned}
\hat{b}_1(r_1) &= \tfrac{1}{2}(c, r_2) + \tfrac{1}{2}(d, t_2'),\\
\hat{b}_2(t_1'') &= \tfrac{1}{2}(c, t_2') + \tfrac{1}{2}(d, r_2),\\
\hat{b}_2(r_2) &= \tfrac{3}{4}(e, r_1) + \tfrac{1}{4}(f, t_1''),\\
\hat{b}_2(t_2') &= \tfrac{1}{4}(e, r_1) + \tfrac{3}{4}(f, t_1'').
\end{aligned}$$

It can be verified that $\hat{\mathcal{T}}$ is indeed non-redundant, and induces the same collection of belief hierarchies as the original type structure $\mathcal{T}$ from Table 1.

Another global question we can answer is whether two different type structures generate the same *collection* of belief hierarchies, or whether the collection of belief hierarchies induced by the first type structure is contained in that of the second structure. This is the type of question which is addressed, for instance, in Friedenberg and Meier (2011). To answer the first question we can first merge the two type structures into one, and subsequently run the *Type Partitioning Procedure.* If the final partitions are such that every equivalence class always contains at least one type from both type structures, then the two structures generate the same collection of belief hiearchies. Otherwise not. Indeed, assume that every equivalence class in the final partitions contains at least one type from each type structure. Then, for every type in the first structure there is a type in the second structure that generates the same belief hierarchy, and *vice versa.* That is, both type structures produce exactly the same collection of belief hierarchies. If this is not the case, that is, if there is an equivalence class that contains only types from one type structure but not from the other, then these type do not have any "counterpart" in the other type structure, and hence the two type structures differ in the collection of belief hierarchies they generate. To answer the second question – that is, whether the set of belief hierarchies of the first structure is contained in that of the second – we look at the final partitions, and see whether every equivalence class contains at least one type from the second structure.

We have collected the insights above in the following corollary.

Corollary 2 (Type Structures). *Consider a multi-agent uncertainty space $\mathcal{X} = (X_i, \Sigma_i)_{i \in I}$, and two finite type structures $\mathcal{T}^1 = (T_i^1, b_i^1)_{i \in I}$ and $\mathcal{T}^2 = (T_i^2, b_i^2)$ for $\mathcal{X}$. Let $(\mathcal{P}_i^\infty)_{i \in I}$ be the final partitions generated by the Type Partitioning Procedure if we first merge the two type structures into one. Then:*

(a) type structure $\mathcal{T}^1$ is redundant, if and only if, there is some agent i and some $P_i \in \mathcal{P}_i^\infty$ such that $|P_i \cap T_i^1| \geq 2$;

(b) the collection of belief hierarchies induced by $\mathcal{T}^1$ is a subset of the collection of belief hierarchies induced by $\mathcal{T}^2$, if and only if, $P_i \cap T_i^2 \neq \emptyset$ for all agents i and all $P_i \in \mathcal{P}_i^\infty$;

(c) type structures $\mathcal{T}^1$ and $\mathcal{T}^2$ induce the same collection of belief hierarchies, if and only if, $P_i \cap T_i^1 \neq \emptyset$ and $P_i \cap T_i^2 \neq \emptyset$ for all agents i and all $P_i \in \mathcal{P}_i^\infty$.

Here, we say that a type structure is redundant if it contains at least two different types that generate the same belief hierarchy.

6. PROOFS

Proof of Theorem 1. We first prove (a) by induction on n.

Induction start. The partition $\mathcal{P}_i^1$ will always be a refinement of $\mathcal{P}_i^0$ since $\mathcal{P}_i^0$ is the trivial partition, by definition.

Inductive step. Let $n \geq 2$, and suppose that $\mathcal{P}_i^{n-1}$ is a refinement of $\mathcal{P}_i^{n-2}$, for all agents i. Consider an agent i, an equivalence class $P_i^n \in \mathcal{P}_i^n$, and two types $t_i, t_i' \in P_i^n$. Then, by equation (2),

$$\begin{aligned}
b_i(t_i)(E_i \times P_{-i}^{n-1}) &= b_i(t_i')(E_i \times P_{-i}^{n-1}) \text{ for all } E_i \in \Sigma_i,\\
&\quad \text{and all } P_{-i}^{n-1} \in \mathcal{P}_{-i}^{n-1}.
\end{aligned}$$

As, by the induction assumption, $\mathcal{P}_j^{n-1}$ is a refinement of $\mathcal{P}_j^{n-2}$ for all $j \neq i$, it follows that $\mathcal{P}_{-i}^{n-1}$ is a refinement of $\mathcal{P}_{-i}^{n-2}$. But then, we conclude from (3) that

$$\begin{aligned}
b_i(t_i)(E_i \times P_{-i}^{n-2}) &= b_i(t_i')(E_i \times P_{-i}^{n-2}) \text{ for all } E_i \in \Sigma_i,\\
&\quad \text{and all} P_{-i}^{n-2} \in \mathcal{P}_{-i}^{n-2},
\end{aligned}$$

which means that t_i and t'_i belong to the same equivalence class in $\mathcal{P}_i^{n-1}$. So, we have shown that every two types that are in the same equivalence class of $\mathcal{P}_i^n$, are also in the same equivalence class of $\mathcal{P}_i^{n-1}$. This, however, implies that $\mathcal{P}_i^n$ is a refinement of $\mathcal{P}_i^{n-1}$, as was to show. By induction on n, property (a) follows.

With property (a) at hand, it is easy to prove property (b). By property (a) we know that for every round $n \geq 1$, and every agent i, the partition $\mathcal{P}_i^n$ is a refinement of $\mathcal{P}_i^{n-1}$. Moreover, for every "active" round n – where the procedure does not terminate yet – the partition $\mathcal{P}_i^n$ must be a *strict* refinement of $\mathcal{P}_i^{n-1}$ for at least one agent i. It may be verified that for every agent i, the number of successive strict refinements cannot be larger than the number of types in T_i. As such, the number of active rounds in the procedure cannot be larger than the number of types in $\cup_{i\in I} T_i$, which is N. This completes the proof. ■

Proof of Theorem 2. We first prove (a) by induction on n.

Induction start. Consider two types $t_i, t'_i \in T_i$. Suppose first that $h_i^1(t_i) = h_i^1(t'_i)$. We show that $t'_i \in \mathcal{P}_i^1(t_i)$. For all $E_i \in \Sigma_i$ and $P_{-i}^0 \in \mathcal{P}_{-i}^0$,

$$\begin{aligned} b_i(t_i)(E_i \times P^0_{-i}) &= b_i(t_i)(E_i \times T_{-i}) \\ &= h_i^1(t_i)(E_i) \\ &= h_i^1(t'_i)(E_i) \\ &= b_i(t'_i)(E_i \times T_{-i}) \\ &= b_i(t'_i)(E_i \times P^0_{-i}), \end{aligned}$$

which indeed implies that $t'_i \in \mathcal{P}_i^1(t_i)$. Here, the first and fifth equality follow from the fact that $\mathcal{P}_{-i}^0$ is the trivial partition on T_{-i}, the second and fourth equality follow from the definition of $h_i^1(t_i)$ and $h_i(t'_i)$, respectively, whereas the third equality follows from the assumption that $h_i^1(t_i) = h_i^1(t'_i)$.

Assume next that $t'_i \in \mathcal{P}_i^1(t_i)$. We show that $h_i^1(t_i) = h_i^1(t'_i)$. For all $E_i \in \Sigma_i$,

$$\begin{aligned} h_i^1(t_i)(E_i) &= b_i(t_i)(E_i \times T_{-i}) = b_i(t'_i)(E_i \times T_{-i}) \\ &= h_i^1(t'_i)(E_i), \end{aligned}$$

which indeed implies that $h_i^1(t_i) = h_i^1(t'_i)$. Here, the first and third equality follow from the definition of $h_i^1(t_i)$ and $h_i^1(t'_i)$, respectively, whereas the second equality follows from the assumption that $t'_i \in \mathcal{P}_i^1(t_i)$ and equation (2).

By the two steps above we may conclude that $h_i^1(t_i) = h_i^1(t'_i)$, if and only if, $t'_i \in \mathcal{P}_i^1(t_i)$, as was to show.

Inductive step. Let $n \geq 2$, and assume that (a) holds for $n-1$ and all agents i. Consider an agent i and two types $t_i, t'_i \in T_i$. Suppose first that $h_i^n(t_i) = h_i^n(t'_i)$. We show that $t'_i \in \mathcal{P}_i^n(t_i)$.

Consider some $P_{-i}^{n-1} \in \mathcal{P}_{-i}^{n-1}$. Take some arbitrary $t_{-i} \in P_{-i}^{n-1}$ and let $h_{-i}^{n-1} = h_{-i}^{n-1}(t_{-i})$. By the induction assumption, P_{-i}^{n-1} contains all type combinations in T_{-i} that induce the same combination of $(n-1)$-th order beliefs as t_{-i}. Remember from Section 2.2 that $T_{-i}[h_{-i}^{n-1}]$ denotes the set of type combinations in T_{-i} that induce h_{-i}^{n-1}. Then, we may conclude that $P_{-i}^{n-1} = T_{-i}[h_{-i}^{n-1}]$. For every $E_i \in \Sigma_i$ we then have that

$$\begin{aligned} b_i(t'_i)(E_i \times P_{-i}^{n-1}) &= b_i(t'_i)(E_i \times T_{-i}[h_{-i}^{n-1}]) \\ &= h_i^n(t'_i)(E_i \times \{h_{-i}^{n-1}\}) \\ &= h_i^n(t_i)(E_i \times \{h_{-i}^{n-1}\}) \\ &= b_i(t_i)(E_i \times T_{-i}[h_{-i}^{n-1}]) \\ &= b_i(t_i)(E_i \times P_{-i}^{n-1}), \end{aligned}$$

which by equation (2) indeed implies that $t'_i \in \mathcal{P}_i^n(t_i)$. Here, the first and fifth equality follows from the insight above that $P_{-i}^{n-1} = T_{-i}[h_{-i}^{n-1}]$, the second and the fourth equality follow from the definition of $h_i^n(t'_i)$ and $h_i^n(t_i)$, whereas the third equality follows from the assumption that $h_i^n(t_i) = h_i^n(t'_i)$.

Suppose next that $t'_i \in \mathcal{P}_i^n(t_i)$. We show that $h_i^n(t_i) = h_i^n(t'_i)$.

Take some arbitrary combination $h_{-i}^{n-1} \in h_{-i}^{n-1}(T_{-i})$ of $(n-1)$-th order beliefs that is obtained by at least one type combination in T_{-i}. By the induction assumption, there must be some $P_{-i}^{n-1} \in \mathcal{P}_{-i}^{n-1}$ such that $P_{-i}^{n-1} = T_{-i}[h_{-i}^{n-1}]$. Then, for every $E_i \in \Sigma_i$,

$$\begin{aligned} h_i^n(t_i)(E_i \times \{h_{-i}^{n-1}\}) &= b_i(t_i)(E_i \times T_{-i}[h_{-i}^{n-1}]) \\ &= b_i(t_i)(E_i \times P_{-i}^{n-1}) \\ &= b_i(t'_i)(E_i \times P_{-i}^{n-1}) \\ &= b_i(t'_i)(E_i \times T_{-i}[h_{-i}^{n-1}]) \\ &= h_i^n(t'_i)(E_i \times \{h_{-i}^{n-1}\}), \end{aligned}$$

which indeed implies that $h_i^n(t_i) = h_i^n(t'_i)$. Here, the third equality follows from the assumption that $t'_i \in \mathcal{P}_i^n(t_i)$ and equation (2), whereas the other equalities follow exactly as above.

By the two steps above we may thus conclude that $h_i^n(t_i) = h_i^n(t'_i)$, if and only if, $t'_i \in \mathcal{P}_i^n(t_i)$. By induction on n, statement (a) follows.

The proof of (b) follows immediately from (a) and property (b) in Theorem 1. This completes the proof. ■

7. REFERENCES

[1] Böge, W. and T.H. Eisele (1979), On solutions of bayesian games, *International Journal of Game Theory* **8**, 193–215.

[2] Brandenburger, A. (2007), The power of paradox: Some recent developments in interactive epistemology, *International Journal of Game Theory* **35,** 465–492.

[3] Brandenburger, A. and E. Dekel (1987), Rationalizability and correlated equilibria, *Econometrica* **55**, 1391–1402.

[4] Dekel, E., D. Fudenberg and S. Morris (2007), Interim correlated rationalizability, *Theoretical Economics* **2,** 15–40.

[5] Dekel, E. and M. Siniscalchi (2013), Epistemic game theory, Chapter prepared for *Handbook of Game Theory.*

[6] Ely, J.C. and M. Pęski (2006), Hierrachies of belief and interim rationalizability, *Theoretical Economics* **1,** 19–65.

[7] Friedenberg, A. and M. Meier (2011), On the relationship between hierarchy and type morphisms, *Economic Theory* **46,** 377–399.

[8] Harsanyi, J.C. (1962), Bargaining in ignorance of the opponent's utility function, *Journal of Conflict Resolution* **6**, 29–38.

[9] Harsanyi, J.C. (1967–1968), Games with incomplete information played by "bayesian" players, I–III, *Management Science* **14**, 159–182, 320–334, 486–502.

[10] Heifetz, A. and W. Kets (2013), Robust multiplicity with a grain of naiveté, Working paper.

[11] Heifetz, A. and D. Samet (1998), Topology-free typology of beliefs, *Journal of Economic Theory* **82,** 324–341.

[12] Mertens, J.-F. and S. Zamir (1985), Formulation of bayesian analysis for games with incomplete information, *International Journal of Game Theory* **14**, 1–29.

[13] Perea, A. (2012), *Epistemic Game Theory: Reasoning and Choice,* Cambridge University Press.

[14] Tan, T. and S.R.C. Werlang (1988), The bayesian foundations of solution concepts of games, *Journal of Economic Theory* **45**, 370–391.

[15] Weinstein, J. and M. Yildiz (2007), Impact of higher-order uncertainty, *Games and economic Behavior* **60,** 200–212.

Rationalization and Robustness in Dynamic Games with Incomplete Information

[Extended Abstract]

Evan Piermont
University of Pittsburgh
4200 5th Avenue
Pittsburgh, PA, USA
ehp5@pitt.edu

Peio Zuazo-Garin
Universitat Rovira i Virgili
Avinguda de la Universitat 1
43204, Reus, Spain
peio.zuazo@urv.cat

Keywords

Dynamic Games, Common Knowledge, Rationalization, Robustness.

1. INTRODUCTION

In the game theoretic environment, there is a clear tension between the strength of a solution concept and its robustness to misspecification. In other words, if the analyst wants his model to be resilient to small errors in the parameters, then he must weaken the predictive power of the model. In fact, this tension can be made formal; under a richness assumption (loosely speaking, for every strategy there is a state such that it is dominant), structure theorems place clear limits to the predictive power of robust solution concepts [8, 6].

Intuitively, a similar tension arises when considering a solution concept and its epistemic demands.[1] To make sharper predictions, the modeler must place more stringent requirements on the structure of the understanding of agents (at least insofar as to adhere to the requirements of the solution concept). Informally, this observation suggests a possible link between the epistemic demands of a solution concept and it's robustness. The first aim of this paper is to formalize this connection. We show that particular notions of robustness can be thought of as epistemic concerns. In particular, we examine a solution concept's robustness to the misspecification of players' beliefs, the underlying space of payoff uncertainty, and to the joint misspecification of both. In each case, we show that reasonable and common notions of robustness can be described entirely though the epistemic characterization of the solution concept.

Most commonly, robustness has been defined with respect to misspecification of player's beliefs; in particular via the upper-hemicontinuity (henceforth, UHC) of the solution concept in question. UHC dictates that if a strategy is ruled out for some type of player, then there is a neighborhood of nearby types for which the strategy is also ruled out. In the absence of UHC, approximations will not suffice; even if a strategy is selected by every successive approximation, it may not be selected in the limit.

Our first result provides the epistemic characterization of UHC, and hence, a direct method of verification of robustness. The UHC of a a solution concept is related to the closedness of the event that characterizes it. Intuitively, that a solution concept is well behaved with respect to approximation of the players types is implied by the fact that the limit of any sequence in the characterizing event is also in the event.

A second aim of this paper is to develop an appropriate notion of robustness to the misspecification of the space of payoff uncertainty (referred to as the *state space*). Although the richness assumption allows for structure theorems that place clear limits on the robustness of solution concepts and provide for generic dominance solvability, it is often an unreasonable demand on the space of uncertainty. To assume richness is to drop all common knowledge assumptions regarding strategic unceratainty.[2] However, in many, if not most, economic situations, there are clear restrictions on payoffs. Without assuming major structural ignorance on the part of the players, the economic restrictions are best modeled by common knowledge assumptions. Moreover, simply embedding the game in a larger state space (and modeling common knowledge as initial common belief) may distort predictions (for example, see section 2.1). This is a problem exclusive to dynamic environments, as in static games, the players' initial (and only) beliefs entirely govern actions. In dynamic environments, however, when players are able to update their beliefs, the state space itself becomes relevant to the analysis.

Penta [6] proposes the robustness notion of *informational invariance*, or, that the predictions of a solution concept do not change when the game is embedded in *any* larger state space. We ague that this is too strong a requirement, as it restricts the solution concept from having any dependence on the state space that is not present in the player's initial beliefs. Instead, we propose the notion of state-space-robustness (s-robustness). To this end, we allow the space of uncertainty that each player considers to become a parameter of the game. We call each players understanding about the state space, and his higher order understanding about his opponents understanding of the state space, his *directory*. A player is described not only by his beliefs (a hierarchy of probability distributions), but also by his directory (a hierarchy of sets of parameters). Then, in analogy to UHC, a solution concept is s-robust if whenever a strategy is ruled out for a directory there is a neighborhood of nearby

[1] Informally, epistemic demand refer to the restrictions placed on players beliefs regarding payoff uncertainty, opponents strategies, and the higher order beliefs over these objects. A more formal, but my no means complete, explanation is found in Section 2.

[2] *Relaxing* common knowledge assumptions can be somehow regarded as *strengthening* common awareness ones.

directories for which the strategy is also ruled out. As in regards to UHC, we provide the epistemic restrictions that coincide with s-robustness.

Lastly, we provide a structure theorem that identifies conditions on directories such that any strict refinement of Extensive Form Rationalizability (EFR) (introduced in [5]) is not UHC. We find conditions that are strictly weaker than the richness assumption in [6] and its static counterpart in [8]. In particular, while we require the existence of some objective rich state space, this state space need not be commonly known. Each player could be described by a directory that does *not* contain dominance states. Indeed, it is possible to obtain the result even in the case where it is common knowledge that no action is ever dominant.

2. UPPER-HEMICONITINUITY: AN EPISTEMIC APPROACH

We consider dynamic environments and our analysis takes place within a formal epistemic framework. Players in a game hold beliefs regarding the state space Θ, the other players' beliefs about the payoff uncertainty, and other players' strategies S_{-i}. Moreover, players hold higher-order beliefs about other players' beliefs about these objects, beliefs about these beliefs and so on.

Following closely the construction due to [2], we formally represent each player i's higher-order conditional beliefs on opponents' choices and the payoff state via conditional belief hierarchies, i.e., (epistemic) types drawn from a *universal* type space we denote by $\mathcal{E}$. Note that players' uncertainty in $\mathcal{E}_{-i} \times S_{-i} \times \Theta$ is not modeled only at the beginning of the game, but also at every history along the possible paths of play. We further assume that beliefs are updated in a Bayesian manner whenever possible.

The epistemic analysis is then performed in the set of *states of the world*, $\Omega = \mathcal{E} \times S \times \Theta$. Each epistemic type, $e_i \in \mathcal{E}_i$, induces a standard type, $\tau_i \in \mathcal{T}_i$ via the canonical quotient map, $q_i : \mathcal{E}_i \to \mathcal{T}_i$.[3] For each standard type, τ_i, we denote by $[q_i = \tau_i]$ the event that player i's standard hierarchy is exactly τ_i. A solution concept can then be characterized by an event contained in this state space.

Definition 1. *An event $E \subseteq \Omega$ **characterizes** the solution concept $\mathcal{S}_i : \mathcal{T}_i \rightrightarrows S_i$ if for all τ_i it holds that $\mathcal{S}_i(\tau_i) = \text{Proj}_{S_i}(E \cap [q_i = \tau_i])$.*

Throughout the paper we focus on two different dynamic solution concepts: *extensive-form rationalizability* (EFR) [5], and *interim sequential rationalizable* (ISR) [6]. Both are generalizations of the static notion of interim correlated rationalizability; the difference being that EFR requires that players place a higher epistemic priority to rationality than ISR. Our focus on ISR is driven by the structure theorem of [6], which states that, under the richness assumption, any strict refinement of ISR is not UHC.

The epistemic characterization of ISR, provided by [6], is the event composed of Rationality (**R**) and Initial Common Belief in Rationality (**ICBR**). **R** states that each player will always choose a strategy that is a best response to his own belief. **ICBR** for player i states that, at the beginning of the game, player i believes players $j \neq i$ are rational, believes that players $j \neq i$ believe players $k \neq j$ are rational, etc. Since players are Bayesian, this implies that at any history reached with positive probability according to player i's initial belief, player i still holds a common belief in **R**.

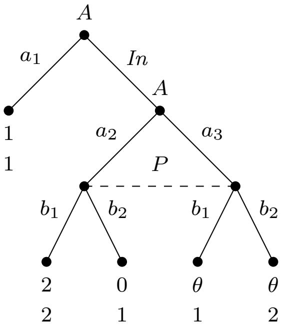


Figure 1: Examples 1 and 2 in [6].

In this paper, we recall the characterization of EFR as the event composed of **R** and Common Strong Belief in Rationality (**CSBR**). Strong Belief in Rationality for player i is the event that player i assigns probability 1 to event **R** at *every* history which is not belief-inconsistent with **R**, or, in other words, whenever **R** has not been falsified by observed history. This way, strong belief captures the essence of *forward induction*, in the sense that beliefs about future behavior are, whenever possible, updated according to observed history and the filter rationality imposes. **CSBR** is then the hierarchical iteration of strong belief in rationality; that is, as stated by Battigalli's *best rationalization principle*, the event that players assign to their opponents the highest degree of strategic sophistication consistent with observed behavior [1]. It is immediate that **ICBR** is implied by **CSBR** and hence EFR is a refinement of ISR. This last observation, together with the structure theorem in [6], suggests a clear tension between rationalization and robustness: under richness, the robustness of predictions appears to be lost as soon as players are assumed to reason according to the highest rationalization principle. The following example illustrates the conflict.

2.1 Alexei and Polina

Example 1 by [6] studies the sequential game depicted in Figure 1, in which Alexei Ivanovich's (player A) utility after history (In, a_3) is represented by unspecified parameter θ. We focus first in the case in which Alexei and Polina Alexandrovna (player P) commonly know that θ equals 0: $\Theta = \{0\}$ and the the corresponding standard type space is $\mathcal{T}^{CK} = \{\tau^{CK}\}$. In this case, strategy (In, a_3) is strictly dominated for Alexei. So, if Polina finds herself at information set $\{(In, a_2), (In, a_3)\}$ (i.e, she is *informed* that Alexei did not choose a_1) *and* believes that Alexei is rational, then she must believe he played (In, a_2). Thus, action b_1 is optimal for her. Now, if Alexei expects Polina to rationalize his choice, he is able to predict choice b_1 in case he plays (In, a_2), and hence, (In, a_2) becomes optimal for him. However, if Polina finds herself at information set $\{(In, a_2), (In, a_3)\}$ *and* does not believe that Alexei is rational, she can rationalize any action. Hence ISR, which does not require Polina to

[3] The standard type space consists of all hierarchies of beliefs over payoff uncertainty, modeled at the null history. It is analogous to the canonical space constructed for static environments by [3]. The map q_i is simply the hierarchical marginalization on this space.

believe Alexei is rational at $\{(In, a_2), (In, a_3)\}$ (if, say, she placed probability 1 on his playing a_1), considers strictly more strategies than EFR, which makes such a requirement. Indeed, ISR$(\tau^{CK}) = \{a_1, (In, a_2)\} \times \{b_1, b_2\}$, and EFR$(\tau^{CK}) = ((In, a_2), b_2)$, as shown in greater detail in [6].

Next, consider the model in which $\Theta = \{0, 3\}$ and there is common certainty that $\theta = 0$, that is $\mathcal{T}^{CC} = \{\tau^{CC}\}$, such that type τ^{CC} assigns probability 1 to payoff state and opponents' type combination $(0, \tau^{CC})$.[4] The analysis in [6] constructs a set of types $\{\tau^m\}_{m\in\mathbb{N}}$ such that $\tau^m \to \tau^{CC}$ and ISR$(\tau^m) = \{(a_1, b_2)\}$ for all m. Now, if EFR(τ^{CC}) = EFR$(\tau^{CK}) = \{((In, a_2), b_2)\}$ this would indeed be a violation of UHC, as EFR is a refinement of ISR, hence, for all types in the sequence $\{\tau^m\}_{m\in\mathbb{N}}$, EFR$(\tau^m) \subseteq$ ISR(τ^m) = $\{(a_1, b_2)\}$.

Notice however, that EFR$(\tau^{CC}) \neq$ EFR(τ^{CK}); the move from common knowledge to common certainty changes the set of EFR strategies. Although Polina must retain his belief that Alexei is rational, even after surprising events, it is *not* the case that she must retain his belief that Alexei places probability 1 on $\theta = 0$. Indeed, consider the following first-order beliefs: Alexei assigns probability 1 to $(b_2, 0)$ when $\theta = 0$, and probability 1 to $(b_2, 3)$ when $\theta = 3$, and Polina assigns probability 1 to $(a_1, 0)$ at the beginning of the game, and probability 1 to $((In, a_3), 3)$ when she observes In. Now, assume that Alexei's second-order belief assigns probability 1 to Polina's first-order beliefs above. Polina's second-order beliefs assign probability 1 to Alexei's first-order belief corresponding to $\theta = 0$ at the beginning of the game, and, when she observes In, assign probability 1 to Alexei's first-order belief corresponding to $\theta = 3$. Keeping this iteration, it is easy to check that we obtain a profile of conditional belief hierarchies that represent **CSBR** and initial common belief in $\theta = 0$ (when $\theta = 0$, something Alexei is informed about). Moreover, the unique best-response is indeed profile (a_1, b_2). Hence, $(a_1, b_2) \in$ EFR(τ^{CC}). The mere fact that updated beliefs may take $\theta = 3$ into account, and that it is commonly known that this is possible, mutes the restriction imposed by the high epistemic priority is given to rationality.

2.2 Characterization

This example illustrates two key issues. The first is that the aforementioned tension between rationalization and robustness is *not* present: EFR does not fail to be UHC in the example. The tension is mitigated since Polina can revise her beliefs about Alexei's actions without losing his belief in Alexei's rationality because there exists some state ($\theta = 3$) that allowed Alexei to be rational (but perhaps with incorrect beliefs) and take the particular action. This is a special case of a more general phenomena: when all common knowledge assumptions are dropped (i.e., under richness) this observation extends to *any* action: if an action could be rationalized by assuming an opponent is irrational, then it can be rationalized by changing only that opponents beliefs, but retaining his rationality. These observations can be made formal, and are best examined thorough the following result:

Proposition 1. *Let $E \subseteq \Omega$ be a closed event such that $E \cap [q_i = \tau_i] \neq \emptyset$ for any standard type τ_i. Then, for player i, the following correspondence is upper-hemicontinuous:*

$$\begin{array}{rccl} \mathcal{S}_i^E : & \mathcal{T}_i & \rightrightarrows & S_i \\ & \tau_i & \to & \mathrm{Proj}_{S_i}\left(R_i \cap E \cap [q_i = \tau_i]\right). \end{array}$$

Proposition 1 provides a general way of verifying if the a solution concept is UHC. Note, by requiring that $E \cap [q_i = \tau_i]$ is non empty, we are implicitly requiring that the solution concept does not place any direct restrictions on the standard type space. In other words, this method of verification only works if the solution concept is well defined for all possible standard types.

Utilizing this result and the characterization of EFR (i.e., by setting $E =$ **CSBR**), it is straightforward to check that the result holds for EFR. This observation, along with the structure theorem in [6], easily delivers the following result:

Corollary 1. EFR *is UHC. Moreover, under richness,* ISR *and* EFR *coincide.*

Since under richness no strict refinement of ISR is UHC, and since EFR is both UHC and a refinement of ISR, the two solution concepts must coincide under richness. Intuitively this is because **CSBR** only induces a binding restriction on player i's conjectures about his opponents strategies under common knowledge assumptions. This is an important distinction, since it implies that the tension between robustness and rationalization would only be assured to exist when rationalization has no effect at all.

3. PERSONAL BASE SPACES OF UNCERTAINTY AND S-ROBUSTNESS

The second issue raised by the example is the importance of considering the state space, and the inability to maintain predictions under embeddings. Indeed, the initial intuition of the example, that EFR is not UHC, is being driven not by a failure of convergence of standard types (since EFR it is in fact UHC), but by the failure to foresee the effects of changing the space of strategic uncertainty. In light of this, we propose a framework where players each have a subjective understanding of the space of payoff uncertainty. This accomplishes two goals simultaneously. First, it allows for the modeler to examine how changes (or mis-specifications) of the true state-space effect predictions. Second, it relaxes the restriction that the state-space is commonly known.

In regards to the second point, it is worth noting that while the richness assumption can be interpreted as relaxing all common knowledge restrictions regarding the payoffs associated with a given strategy, it still imposes, rather restrictively, that all players commonly know that all such common knowledge restrictions are relaxed. In other words, that all players commonly know that it is not common knowledge that any action is not dominant. Hence, by allowing the space of uncertainty to be subjective, we relax the imposition that the state space is commonly known.

To do this we define each players *directory*. Beginning with some fixed, objective state space Θ^0, a directory for player i assigns the subset of Θ^0 that he understands to be the true space of parameters, the subset he understands each of his opponents to understand to be the true space, and so on. Notationally, let $\mathcal{K}(\Theta^0)$ be the set of all non-empty compact subsets of Θ^0, let I be the set of players, and for each $n \in \mathbb{N}$ let $\Lambda^n = I^n$; finally set $\Lambda = \bigcup_{n\in\mathbb{N}} \Lambda^n$.

[4] Penta actually considers a slightly different model in which P1 has private information, but the analysis is undistorted.

Definition 2. *For each player $i \in I$, we say that $\Theta_i : \Lambda \to \mathcal{K}(\Theta^0)$ is a* ***directory for player i****, if for any $n \in \mathbb{N}$ and any $\lambda \in \Lambda^n$:*

D1. If $\lambda_1 = i$, then $\Theta_i(\lambda) = \Theta_i\left((\lambda_k)_{k=2}^n\right)$.

D2. If $\lambda_k = \lambda_{k+1}$ for some $k < n$, then

$$\Theta_i(\lambda) = \Theta_i\left((\lambda_1, \dots, \lambda_{k-1}, \lambda_{k+1}, \dots, \lambda_n)\right).$$

Let $\mathcal{D}(\Theta^0)$ be the set of all directories defined over Θ^0.

If $\Theta_i(j) = \Theta_1$ then player i only takes into account Θ_1 when conjecturing about j's first order beliefs. Likewise, if $\Theta_i(j,k) = \Theta_2$, then player i only takes into account Θ_2 when conjecturing about j's second order beliefs about k's first order beliefs. Under our interpretation, D1 states that each player correctly understands his own understanding. Then, D1 implies that $\Theta_i(i,j) = \Theta_i(i,i,j) = \Theta_1$, etc. Similarly, D2 dictates that in each player's mind all other players understand the restriction imposed by D1. Hence, player 1's understands that player 2 understands his own understanding. Hence, $\Theta_i(i,j,j,j) = \Theta_i(i,j,j) = \Theta_1$, too. Note that each player i's directory Θ_i induces a directory for all of his opponents in his mind, namely $\Theta_{j|i}$, where $\Theta_{j|i}(\lambda) = \Theta_i(j,\lambda)$ for any $\lambda \in \Lambda$ and any opponent j. As usual, we denote $\Theta_{-i|i} = \prod_{j \neq i} \Theta_{j|i}$.

Each directory, Θ_i, induces a type space, $\mathcal{T}_i^{\Theta_i}$, in a natural way. $\mathcal{T}_i^{\Theta_i}$ is the set of types (drawn from the universal space over Θ^0) whose beliefs are concentrated on the appropriate set assigned by the directory. Each player can be described by a pair (Θ_i, τ_i), where $\tau_i \in \mathcal{T}_i^{\Theta_i}$.

By directly incorporating each players understanding of the state space, it becomes easy to define s-robustness.

Definition 3. *Let $\mathcal{G}$ be a game with incomplete information. Then, we say that solution concept $\mathcal{S}_i : \mathrm{Graph}(\mathcal{T}_i^{(\cdot)}) \rightrightarrows S_i$ is* ***s-robust****, if for any player i, standard hierarchy τ_i, and sequence $(\Theta_i^n, s_i^n)_{n\in\mathbb{N}} \subseteq \mathcal{D}(\Theta^0) \times S_i$ such that:*[5]

(i) $\tau_i \in \mathcal{T}_i^{\Theta_i^n}$ for all n,

(ii) $s_i^n \in \mathcal{S}_i(\Theta_i^n, \tau_i)$ for all n,

(iii) $\lim_{n\to\infty} \Theta_i^n = \Theta_i$, and,

(iv) $\lim_{n\to\infty} s_i^n = s_i$,

then $s_i \in \mathcal{S}_i(\Theta_i, \tau_i)$.

Just as UHC states that the solution concept should be consistent in the limit of successive approximations of player's true type, s-robustenss imparts the same requirement on approximation of the player's true understanding of the true state-space. One can model the situation where some state space, Θ_1, is commonly known by setting $\Theta_i(\lambda) = \Theta_1$ for all i and λ. It is within this specific case that we can model embeddings of the game and get a more direct comparison to informational invariance [6]. S-robustness is a significantly weaker requirement than informational invariance, as it only requires the solution concept to be resilient to small misspecifications. It should therefore not be too surprising that EFR is also s-robust.

Proposition 2. *EFR is s-robust.*

This result, like its analog for UHC, is proven by examining the epistemic demands of the robustness criterion. However, this involves a large notational burden, and so, is left out of this abstract.

3.1 A Structure Theorem

Following the literature on robustness, we provide a structure theorem. While the motivation is similar, the notion of directories raises a new question: is the generic uniqueness of dominance solvability –being driven by relaxation of common knowledge conditions on strategic uncertainty– still present under the relaxation of common knowledge conditions on payoff uncertainty. That is, can we still obtain a structure theorem if we allow players to understand the state space as not being rich.

We provide a (partial) affirmative answer. We find strictly weaker conditions under which a structure theorem still obtains.

Assumption 1 (Objective Richness). *Θ^0 satisfies the Richness Condition.*[6]

Although we are assuming that the objective space of uncertainty is sufficiently rich, Assumption 1 contains no common knowledge restrictions. This is the direct result of the disentangling of the objective space of uncertainty with players personal base spaces, as given by the directory.

Definition 4. *We say that $\Theta \in \mathcal{K}(\Theta^0)$ has* ***strongly generic payoffs*** *if for any strategy $s_i \in S_i$, there is a state $\theta \in \Theta$, and a profile of opponents strategies $s_{-i} \in S_{-i}$, such that s_i is a strict best response to s_{-i} at θ.*

Assumption 2 (Common Knowledge of Genericity). *For all i and $\lambda \in \Lambda$, $\Theta_i(\lambda)$ has strongly generic payoffs.*

Assumption 2 is, like richness, a common knowledge restriction on the space of uncertainty. However, it is a significantly weaker restriction. The relationship is as follows: richness dictates that it is not commonly known that any action is not dominant, whereas genericity dictates that it not commonly known that any action is dominant. Indeed, Assumptions 1 and 2 allow for the situation where it is commonly known that *no* action is dominant. For example, a simple "Battle of the Sexes" game with no payoff uncertainty would satisfy Assumption 2 but clearly fail to satisfy richness. Moreover, since richness implies genericity, if Θ^0 is rich and commonly known (i.e., as in previous literature) then Assumptions 1 and 2 are both implied.

Proposition 3. *Under assumptions 1 and 2, any strict refinement of EFR is not UHC. Moreover, the set*

$$\{(\Theta_i, \tau_i) : |\mathrm{EFR}_i(\Theta_i, \tau_i)| = 1\}$$

is open and dense in $\mathrm{Graph}(\mathcal{T}_i^{(\cdot)})$.

Proposition 3 shows that richness need not be commonly known to arrive at generic dominance solvability. The result relies on the ability to construct a sequence of directories that impose richness at increasingly high order understandings; then using Assumption 2 to cascade the effect

[5] In the paper, we formally define the topology on $\mathcal{D}(\Theta^0)$. Intuitively, it is the product topology generated by sequences of $\mathcal{K}(\Theta^0)$, itself endowed with the Hausdorff metric.

[6] As defined in [6].

down without losing convergence. It is worth pointing out that there has been a recent interest in providing structure theorems in the absence of richness [4, 7]. However, these papers focus on the identifying which strategies can be selected when richness is not assumed, rather than weakening the structural conditions on the space of uncertainty while retaining that all rationalizable actions can be selected.

4. CONCLUSION

In this paper we show a formal connection between the epistemic characterization of a solution concept and its robustness to the misspecification of parameters. This provides both an important conceptual link and a direct method for checking robustness when the epistemic characterization is known. We use this result to show that EFR is UHC. We also present a new framework that relaxes the common knowledge restrictions regarding the *space* of payoff parameters. Then, we propose a new type of robustness, s-robustness, to modeling errors of the player understanding of the space of uncertainty, which is of particular importance in dynamic environments. We characterize this notion through our epistemic framework and show that EFR is also s-robust. Finally, we provide a structure theorem for EFR with personal spaces of uncertainty that shows that no common knowledge assumptions regarding the existence of dominance states are required to achieve generic dominance solvability.

5. REFERENCES

[1] P. Battigalli. “Strategic rationality order and the best rationalization principle”. *Games and Economic Behavior*, 13:178–200, 1996.

[2] P. Battigalli and M. Siniscalchi. “Hierarchies of conditional beliefs and interactive epistemology in dynamic games”. *Journal of Economic Theory,*, 88:188–230, 1999.

[3] A. Brandenburger and E. Dekel. “Hierarchies of beliefs and common knowledge”. *Journal of Economic Theory*, 59:189–198, 1993.

[4] Y. C. Chen, S. Takahashi, and S. Xiong. The weinstein-yildiz selection and robust predictions with arbitrary payoff uncertainty. 2014.

[5] D. G. Pearce. “Rationalizable strategic behavior and the problem of perfection”. *Econometrica*, 52:1029–1050, 1984.

[6] A. Penta. “Higher order uncertainty and information: static and dynamic games”. *Econometrica*, 80:631–660, 2012.

[7] A. Penta. “On the structure of rationalizability for arbitrary spaces of uncertainty”. *Theoretical Economics*, 8:405–430, 2013.

[8] J. Weinstein and M. Yildiz. “A structure theorem for rationalizability with application to robust predictions of refinements”. *Econometrica*, 75:365–400, 2007.

Parameterized Complexity Results for a Model of Theory of Mind based on Dynamic Epistemic Logic[*]

Iris van de Pol [†]
Institute for Logic, Language, and Computation
University of Amsterdam
i.p.a.vandepol@uva.com

Iris van Rooij
Donders Institute for Brain, Cognition, and Behaviour
Radboud University
i.vanrooij@donders.ru.nl

Jakub Szymanik [‡]
Institute for Logic, Language, and Computation
University of Amsterdam
j.k.szymanik@uva.nl

ABSTRACT

In this paper we introduce a computational-level model of theory of mind (ToM) based on dynamic epistemic logic (DEL), and we analyze its computational complexity. The model is a special case of DEL model checking. We provide a parameterized complexity analysis, considering several aspects of DEL (e.g., number of agents, size of preconditions, etc.) as parameters. We show that model checking for DEL is PSPACE-hard, also when restricted to single-pointed models and S5 relations, thereby solving an open problem in the literature. Our approach is aimed at formalizing current intractability claims in the cognitive science literature regarding computational models of ToM.

Categories and Subject Descriptors

F.1.3 [**Theory of Computation**]: Complexity Measures & Classes—*Reducibility and Completeness*; F.4.1 [**Mathematical Logic and Formal Languages**]: Mathematical Logic—*Modal Logic*; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Modal Logic*

General Terms

Theory

Keywords

Theory of mind; dynamic epistemic logic; computational complexity; parameterized complexity; computational-level model

1. INTRODUCTION

Imagine that you are in love. You find yourself at your desk, but you cannot stop your mind from wandering off. What is she thinking about right now? And more importantly, is she thinking about you and does she know that you are thinking about her? Reasoning about other people's knowledge, belief and desires, we do it all the time. For instance, in trying to conquer the love of one's life, to stay one step ahead of one's enemies, or when we lose our friend in a crowded place and we find them by imagining where they would look for us. This capacity is known as theory of mind (ToM) and it is widely studied in various fields (see, e.g., [8, 11, 23, 34, 36, 38, 47, 48]).

We seem to use ToM on a daily basis and many cognitive scientists consider it to be ubiquitous in social interaction [1]. At the same time, however, it is also widely believed that computational cognitive models of ToM are intractable, i.e., that ToM involves solving problems that humans are not capable of solving (cf. [1, 27, 31, 50]). This seems to imply a contradiction between theory and practice: on the one hand we seem to be capable of ToM, while on the other hand, our theories tell us that this is impossible. Dissolving this paradox is a critical step in enhancing theoretical understanding of ToM.

The question arises what it means for a computational-level model[1] of cognition to be intractable. When looking more closely at these intractability claims regarding ToM, it is not clear what these researchers mean exactly, nor whether they mean the same thing. In theoretical computer science and logic there are a variety of tools to make precise claims about the level of complexity of a certain problem. In cognitive science, however, this is a different story. With the exception of a few researchers, cognitive scientists do not tend to specify formally what it means for a theory to be intractable. This makes it often very difficult to assess the validity of the various claims in the literature about which theories are tractable and which are not.

In this paper we adopt the *Tractable-cognition thesis* (see [42]) that states that people have limited resources for cognitive processing and human cognitive capacities are confined to those that can be realized using a realistic amount of time.[2] More specifically we adopt the *FPT-*

[*] This research has been carried out in the context of the first author's master's thesis [37].

[†] Supported by Gravitation Grant 024.001.006 of the Language in Interaction Consortium from the Netherlands Organization for Scientific Research.

[‡] Supported by the Netherlands Organisation for Scientific Research Veni Grant NWO-639-021-232.

[1] In cognitive science, often Marr's [33] tri-level distinction between computational-level ("what is the nature of the problem being solved?"), algorithmic-level ("what is the algorithm used for solving the problem?"), and implementational-level ("how is the algorithm physically realized?") is used to distinguish different levels of computational cognitive explanations. In this paper, we will focus on computational-level models of ToM and their computational complexity.

[2] There is general consensus in the cognitive science community that computational intractability is a undesirable feature of cognitive computational models, putting the cognitive plausibility of such models into question [13, 24, 26, 42, 46]. There are diverging opinions about how cognitive

cognition thesis [42] that states that computationally plausible computational-level cognitive theories are limited to the class of input-output mappings that are fixed-parameter tractable for one or more input-parameters that can be assumed to be small in practice. To be able to make more precise claims about the (in)tractability of ToM we introduce a computational-level model of ToM based on dynamic epistemic logic (DEL), and we analyze its computational complexity. The model we present is a special case of DEL model checking. Here we include an informal description of the model.[3] The kind of situation that we want to be able to model, is that of an observer that observes one or more agents in an initial situation. The observer then witnesses actions that change the situation and the observer updates their knowledge about the mental states of the agents in the new situation. Such a set up is often found in experimental tasks, where subjects are asked to reason about the mental states of agents in a situation that they are presented.

DBU (informal) – Dynamic Belief Update
Instance: A representation of an initial situation, a sequence of actions – observed by an observer – and a (belief) statement φ of interest.
Question: Is the (belief) statement φ true in the situation resulting from the initial situation and the observed actions?

We prove that DBU is PSPACE-complete. PSPACE-completeness was already shown by Aucher and Schwarzentruber [3] for DEL model checking in general. They considered unrestricted relations and multi-pointed event models. Since their proof does not hold for the special case of DEL model checking that we consider, we propose an alternative proof. Our proof solves positively the open question in [3] whether model checking for DEL restricted to S5 relations and single-pointed models is PSPACE-complete. Bolander, Jensen and Schwarzentruber [10] independently considered an almost identical special case of DEL model checking (there called the plan verification problem). They also prove PSPACE-completeness for the case restricted to single-pointed models, but their proof does not settle whether hardness holds even when the problem is restricted to S5 models.

Furthermore, we investigate how the different aspects (or parameters, see Table 1) of our model influence its complexity. We prove that for most combinations of parameters DBU is fp-intractable and for one case we prove fp-tractability. See Figure 2 for an overview of the results.

Besides the parameterized complexity results for DEL model checking that we present, the main conceptual contribution of this paper is that it bridges cognitive science and logic, by using DEL to model ToM (cf. [28, 47]). By doing so, the paper provides the means to make more precise statements about the (in)tractability of ToM.

The paper is structured as follows. In Section 2 we introduce basic definitions from dynamic epistemic logic and parameterized complexity theory. Then, in Section 3 we introduce a formal description of our computational-level model and we discuss the particular choices that we make. Next, in Section 4 we present our (parameterized) complexity results. Finally, in Section 5 we discuss the implications of our results for the understanding of ToM.

2. PRELIMINARIES

2.1 Dynamic Epistemic Logic

Dynamic epistemic logic is a particular kind of modal logic (see [16, 6]), where the modal operators are interpreted in terms of belief or knowledge. First, we define epistemic models, which are Kripke models with an accessibility relation for every agent $a \in \mathcal{A}$, instead of just one accessibility relation.

Definition 2.1 (Epistemic model). *Given a finite set $\mathcal{A}$ of agents and a finite set P of propositions, an epistemic model is a tuple (W, R, V) where*

- *W is a non-empty set of worlds;*
- *R is a function that assigns to every agent $a \in \mathcal{A}$ a binary relation R_a on W; and*
- *V is a valuation function from $W \times P$ into $\{0, 1\}$.*

The accessibility relations R_a can be read as follows: for worlds $w, v \in W$, wR_av means "in world w, agent a considers world v possible."

Definition 2.2 ((Multi and single-)pointed epistemic model). *A pair (M, W_d) consisting of an epistemic model $M = (W, R, V)$ and a non-empty set of designated worlds $W_d \subseteq W$ is called a pointed epistemic model. A pair (M, W_d) is called a single-pointed model when W_d is a singleton, and a multi-pointed epistemic model when $|W_d| > 1$. By a slight abuse of notation, for $(M, \{w\})$, we also write (M, w).*

We consider the usual restrictions on relations in epistemic models and event models, such as KD45 and S5 (see [16]). In KD45 models, all relations are transitive, Euclidean and serial, and in S5 models all relations are transitive, reflexive and symmetric.

We define the following language for epistemic models. We use the modal belief operator B, where for each agent $a \in \mathcal{A}$, $B_a\varphi$ is interpreted as "agent a believes (that) φ".

Definition 2.3 (Epistemic language). *The language $\mathcal{L}_B$ over $\mathcal{A}$ and P is given by the following definition, where a ranges over $\mathcal{A}$ and p over P:*

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid B_a\varphi.$$

We will use the following standard abbreviations, $\top := p \vee \neg p, \bot := \neg\top, \varphi \vee \psi := \neg(\neg\varphi \wedge \neg\psi)$, $\varphi \to \psi := \neg\varphi \vee \psi$, $\hat{B_a} := \neg B_a \neg\varphi$.

The semantics for this language is defined as follows.

Definition 2.4 (Truth in a (single-pointed) epistemic model). *Let $M = (W, R, V)$ be an epistemic model, $w \in W$, $a \in \mathcal{A}$, and $\varphi, \psi \in \mathcal{L}_B$. We define $M, w \models \varphi$ inductively as follows:*

science should deal with this issue (see, e.g., [12, 26, 41, 43]). It is beyond the scope of this paper to discuss this in detail. In this paper we adopt the parameterized complexity approach as described in [42].

[3] We pose the model in the form of a decision problem, as this is convenient for purposes of our complexity analysis. Even though ToM may be more intuitively modeled by a search problem, the complexity of the decision problem gives us lower bounds on the complexity of such a search problem, and therefore suffices for the purposes of our paper.

$$\begin{array}{lll} M, w \models p & \textit{iff} & V(w,p) = 1 \\ M, w \models \neg\varphi & \textit{iff} & \textit{not } M, w \models \varphi \\ M, w \models (\varphi \wedge \psi) & \textit{iff} & M, w \models \varphi \textit{ and } M, w \models \psi \\ M, w \models B_a\varphi & \textit{iff} & \textit{for all } v \textit{ with } wR_av\textit{: } M, v \models \varphi \end{array}$$

When $M, w \models \varphi$, we say that φ is true in w or φ is satisfied in w.

Definition 2.5 (Truth in a multi-pointed epistemic model). *Let (M, W_d) be a multi-pointed epistemic model, $a \in \mathcal{A}$, and $\varphi \in \mathcal{L}_B$. $M, W_d \models \varphi$ is defined as follows:*

$$M, W_d \models \varphi \quad \textit{iff} \quad M, w \models \varphi \textit{ for all } w \in W_d$$

Next we define event models.

Definition 2.6 (Event model). *An event model is a tuple $\mathcal{E} = (E, Q, pre, post)$, where E is a non-empty finite set of events; Q is a function that assigns to every agent $a \in \mathcal{A}$ a binary relation R_a on W; pre is a function from E into $\mathcal{L}_B$ that assigns to each event a precondition, which can be any formula in $\mathcal{L}_B$; and post is a function from E into $\mathcal{L}_B$ that assigns to each event a postcondition. Postconditions are conjunctions of propositions and their negations (including $\top$ and $\bot$).*

Definition 2.7 ((Multi and single-)pointed event model / action). *A pair $(\mathcal{E}, E_d)$ consisting of an event model $\mathcal{E} = (E, Q, pre, post)$ and a non-empty set of designated events $E_d \subseteq E$ is called a pointed event model. A pair $(\mathcal{E}, E_d)$ is called a single-pointed event model when E_d is a singleton, and a multi-pointed event model when $|E_d| > 1$. We will also refer to $(\mathcal{E}, E_d)$ as an action.*

We define the notion of a product update, that is used to update epistemic models with actions [4].

Definition 2.8 (Product update). *The product update of the state (M, W_d) with the action $(\mathcal{E}, E_d)$ is defined as the state $(M, W_d) \otimes (\mathcal{E}, E_d) = ((W', R', V'), W'_d)$ where*

- $W' = \{(w, e) \in W \times E \ ; \ M, w \models pre(e)\}$;
- $R'_a = \{((w, e), (v, f)) \in W' \times W' \ ; \ wR_av \textit{ and } eQ_af\}$;
- $V'(p) = 1$ *iff either* $(M, w \models p$ *and* $post(e) \not\models \neg p)$ *or* $post(e) \models p$*; and*
- $W'_d = \{(w, e) \in W' \ ; \ w \in W_d \textit{ and } e \in E_d\}$.

Finally, we define when actions are applicable in a state.

Definition 2.9 (Applicability). *An action $(\mathcal{E}, E_d)$ is applicable in state (M, W_d) if there is some $e \in E_d$ and some $w \in W_d$ such that $M, w \models pre(e)$. We define applicability for a sequence of actions inductively. The empty sequence, consisting of no actions, is always applicable. A sequence $a_1, \ldots, a_k$ of actions is applicable in a state (M, W_d) if (1) the sequence $a_1, \ldots, a_{k-1}$ is applicable in (M, W_d) and (2) the action a_k is applicable in the state $(M, W_d) \otimes a_1 \otimes \cdots \otimes a_{k-1}$.*

2.2 Parameterized Complexity Theory

We introduce some basic concepts of parameterized complexity theory. For a more detailed introduction we refer to textbooks on the topic [17, 18, 22, 35].

Definition 2.10 (Parameterized problem). *Let Σ be a finite alphabet. A* parameterized problem *L (over Σ) is a subset of $\Sigma^* \times \mathbb{N}$. For an* instance *$(x, k)$, we call x the* main part *and k the* parameter*.*

The complexity class FPT, which stands for fixed-parameter tractable, is the direct analogue of the class P in classical complexity. Problems in this class are considered efficiently solvable, because the non-polynomial-time complexity inherent in the problem is confined to the parameter and in effect the problem is efficiently solvable even for large input sizes, provided that the value of the parameter is relatively small.

Definition 2.11 (Fixed-parameter tractable / the class FPT). *Let Σ be a finite alphabet.*

1. *An algorithm* A *with input $(x, k) \in \Sigma \times \mathbb{N}$ runs in* fpt-time *if there exists a computable function f and a polynomial p such that for all $(x, k) \in \Sigma \times \mathbb{N}$, the running time of* A *on (x, k) is at most*

   $$f(k) \cdot p(|x|).$$

   Algorithms that run in fpt-time are called fpt-algorithms*.*

2. *A parameterized problem L is* fixed-parameter tractable *if there is an fpt-algorithm that decides L.* FPT *denotes the class of all fixed-parameter tractable problems.*

Similarly to classical complexity, parameterized complexity also offers a hardness framework to give evidence that (parameterized) problems are not fixed-parameter tractable. The following notion of reductions plays an important role in this framework.

Definition 2.12 (Fpt-reduction). *Let $L \subseteq \Sigma \times \mathbb{N}$ and $L' \subseteq \Sigma' \times \mathbb{N}$ be two parameterized problems. An* fpt-reduction *from L to L' is a mapping $R : \Sigma \times \mathbb{N} \to \Sigma' \times \mathbb{N}$ from instances of L to instances of L' such that there is a computable function $g : \mathbb{N} \to \mathbb{N}$ such that for all $(x, k) \in \Sigma \times \mathbb{N}$:*

1. *$(x', k') = R(x, k)$ is a yes-instance of L' if and only if (x, k) is a yes-instance of L;*
2. *R is computable in fpt-time; and*
3. *$k' \leq g(k)$.*

Another important part of the hardness framework is the parameterized intractability class W[1]. To characterize this class, we consider the following parameterized problem.

$\{k\}$-WSat[2CNF]
Instance: A 2CNF propositional formula φ and an integer k.
Parameter: k.
Question: Is there an assignment $\alpha : var(\varphi) \to \{0, 1\}$, that sets k variables in $var(\varphi)$ to true, that satisfies φ?

The class W[1] consists of all parameterized problems that can be fpt-reduced to $\{k\}$-WSat[2CNF]. A parameterized problem is hard for W[1] if all problems in W[1] can be fpt-reduced to it. It is widely believed that W[1]-hard problems are not fixed-parameter tractable [18]. Another parameterized intractability class, that can be used in a similar way, is the class para-NP. The class para-NP consists of all parameterized problems that can be solved by a nondeterministic fpt-algorithm. To show para-NP-hardness, it suffices to show that DBU is NP-hard for a constant value of the parameters [21]. Problems that are para-NP-hard are not fixed-parameter tractable, unless P = NP [22, Theorem 2.14].

3. COMPUTATIONAL-LEVEL MODEL OF THEORY OF MIND

Next we present a formal description of our computational-level model. Our aim is to capture, in a qualitative way, the kind of reasoning that is necessary to be able to engage in ToM. Arguably, the essence of ToM is the attribution of mental states to another person, based on observed behavior, and to predict and explain this behavior in terms of those mental states. The aspect of ToM that we aim to formalize with our model is the attribution of mental states. There is a wide range of different kinds of mental states such as epistemic, emotional and motivational states. In our model we focus on epistemic states, in particular on belief.

To be cognitively plausible, our model needs to be able to capture a wide range of (dynamic) situations, where all kinds of actions can occur, not just actions that change beliefs (epistemic actions), but also actions that change the state of the world (ontic actions). This is why, following Bolander and Andersen [9], we use postconditions in the product update of DEL (in addition to preconditions).

Furthermore, we want to model the (internal) perspective of the observer (on the situation). Therefore, the god perspective, also called the perfect external approach by Aucher [2] – that is inherent to single-pointed epistemic models – will not suffice for all cases that we want to be able to model. This perfect external approach supposes that the modeler is an omniscient observer that is perfectly aware of the actual state of the world and the epistemic situation (what is going on in the minds of the agents). The cognitively plausible observers that we are interested in here will not have infallible knowledge in many situations. They are often not able to distinguish the actual world from other possible worlds, because they are uncertain about the facts in the world and the mental states of the agent(s) that they observe. That is why, again following Bolander and Andersen [9], we allow for multi-pointed epistemic models (in addition to single-pointed models), which can model the uncertainty of an observer, by representing their perspective as a set of worlds. How to represent the internal or fallible perspective of an agent in epistemic models is a conceptual problem that has not been settled yet in the DEL-literature. There have been several proposals to deal with this (see, e.g., [2, 15, 25]).

Also, since we do not assume that agents are perfectly knowledgeable, we allow the possibility of modeling false beliefs of the observers and agents, by using KD45 models (rather than S5 models). Even though KD45 models present an idealized form of belief (with perfect introspection and logical omniscience), we argue that at least to some extent they are cognitively plausible, and that therefore, for the purpose of this paper, it suffices to focus on KD45 models. Our complexity results (which we present in the next section) do not depend on this choice; they hold for DBU restricted to KD45 models and restricted to S5 models, and also for the unrestricted case.

We define our computational-level model of ToM as follows.

DBU (formal) – Dynamic Belief Update
Instance: A set of propositions P, and set of Agents $\mathcal{A}$. An initial state s_o, where $s_o = ((W, V, R), W_d)$ is a pointed epistemic model. An applicable sequence of actions $a_1, ..., a_k$, where $a_j = ((E, Q, pre, post), E_d)$ is a pointed event model. A formula $\varphi \in \mathcal{L}_B$.
Question: Does $s_o \otimes a_1 \otimes ... \otimes a_k \models \varphi$?

The model can be naturally used to formalize ToM tasks that are employed in psychological experiments. The classical ToM task that is used by (developmental) psychologists is the false belief task [5, 49]. The DEL-based formalization of the false belief task by Bolander [8] can be seen as an instance of DBU. For more details on how DBU can be used to model ToM tasks, we refer to [37].

4. COMPLEXITY RESULTS

4.1 PSPACE-completeness

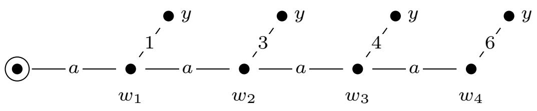
We show that DBU is PSPACE-complete. For this, we consider the decision problem TQBF. This problem is PSPACE-complete [45].

TQBF
Instance: A quantified Boolean formula $\varphi = Q_1x_1Q_2x_2 \dots Q_mx_m.\psi$.
Question: Is φ true?

Theorem 1. DBU *is* PSPACE-*hard.*

Proof. To show PSPACE-hardness we specify a polynomial-time reduction R from TQBF to DBU. Let ψ be a Boolean formula. First, we sketch the general idea behind the reduction. We use the reduction to list all possible assignments to $var(\psi)$. To do this we use groups of worlds (which are R_a-equivalence classes) to represent particular truth assignments. Each group consists of a string of worlds that are fully connected by equivalence relation R_a. Except for the first world in the string, all worlds represent a true variable x_i (under a particular assignment).

We give an example of such a group of worlds that represents assignment $\alpha = \{x_1 \mapsto \text{T}, x_2 \mapsto \text{F}, x_3 \mapsto \text{T}, x_4 \mapsto \text{T}, x_5 \mapsto \text{F}, x_6 \mapsto \text{T}\}$. Each world has a reflexive loop for every agent, which we leave out for the sake of presentation. More generally, in all our drawings we replace each relation R_a with a minimal R'_a whose transitive reflexive closure is equal to R_a. ⦿ marks the designated world. Since all relations are reflexive, we draw relations as lines (leaving out arrows at the end).



We refer to worlds $w_1, \dots, w_4$ as the *bottom worlds* of this group. If a bottom world has an R_i relation to a world that makes proposition y true, we say that it represents variable x_i.

The reduction makes sure that in the final updated model (the model that results from updating the initial state with the actions – which are specified by the reduction) each possible truth assignment to the variables in ψ will be represented by a group of worlds. Between the different groups, there are no R_a-relations (only R_i-relations for $1 \leq i \leq m$). By 'jumping' from one group (representing a particular truth assignment) to another group with relation R_i, the truth value of variable x_i can be set to true or false. We can now translate a quantified Boolean formula into a corresponding formula of $\mathcal{L}_B$ by mapping every universal quantifier Q_i to B_i and every existential quantifier Q_j to $\hat{B}_j$.

To illustrate how this reduction works, we give an example. Figure 1 shows the final updated model for a quantified Boolean formula with variables x_1 and x_2. In this model there are four groups of worlds: $\{w_1, w_2, w_3\}$, $\{w_4, w_5\}$, $\{w_6, w_7\}$ and $\{w_8\}$. Worlds $w_1, \ldots, w_8$ are what we refer to as the bottom worlds. The gray worlds and edges can be considered a byproduct of the reduction; they have no particular function.

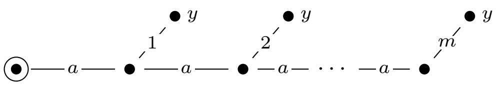
We represent variable x_1 by $\hat{B}_1 y$ and variable x_2 by $\hat{B}_2 y$. Then, in the model above, checking whether $\exists x_1 \forall x_2 . x_1 \vee x_2$ is true can be done by checking whether formula $\hat{B}_1 B_2(\hat{B}_a \hat{B}_1 y \vee \hat{B}a\hat{B}_2 y)$ is true, which is indeed the case. Also, checking whether $\forall x_1 \forall x_2 . x_1 \vee x_2$ is true can be done by checking whether $B_1 B_2(\hat{B}_a \hat{B}_1 y \vee \hat{B}a\hat{B}_2 y)$ is true, which is not the case.

Now, we continue with the formal details. Let $\varphi = Q_1 x_1 \ldots Q_m x_m . \psi$ be a quantified Boolean formula with quantifiers $Q_1, \ldots, Q_m$ and $var(\psi) = \{x_1, \ldots, x_m\}$. We define the following polynomial-time computable mappings. For $1 \leq i \leq m$, let $[x_i] = \hat{B}_i y$, and

$$[Q_i] = \begin{cases} B_i & \text{if } Q_i = \forall \\ \hat{B}_i & \text{if } Q_i = \exists. \end{cases}$$

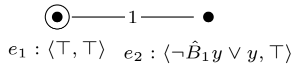
Formula $[\psi]$ is the adaptation of formula ψ where every occurrence of x_i in ψ is replaced by $\hat{B}_a[x_i]$. Then $[\varphi] = [Q_1] \ldots [Q_m][\psi]$. We formally specify the reduction R. We let $R(\varphi) = (P, \mathcal{A}, s_0, a_1, \ldots, a_m, [\varphi])$, where:

- $P = \{y\}$, $\mathcal{A} = \{a, 1, \ldots, m\}$
- $s_0 =$



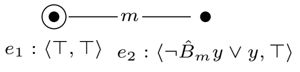
All relations in $s_0, a_1, \ldots, a_m$ are equivalence relations. Note that all worlds in $s_0, a_1, \ldots, a_m$ have reflexive loops for all agents. We omit all reflexive loops for the sake of readability.

- $a_1 =$


1
$e_1 : \langle \top, \top \rangle$
$e_2 : \langle \neg \hat{B}_1 y \vee y, \top \rangle$

$\vdots$

- $a_m =$


m
$e_1 : \langle \top, \top \rangle$
$e_2 : \langle \neg \hat{B}_m y \vee y, \top \rangle$

We show that $\varphi \in$ TQBF if and only if $R(\varphi) \in$ DBU. We prove that for all $1 \leq i \leq m+1$ the following claim holds. For any assignment α to the variables $x_1, \ldots, x_{i-1}$ and any bottom world w of a group that agrees with α, the formula $Q_i x_i \ldots Q_m x_m . \psi$ is true under α if and only if $[Q_i] \ldots [Q_m][\psi]$ is true in world w. In the case for $i = m+1$, this refers to the formula $[\psi]$.

We start with the case for $i = m+1$. We show that the claim holds. Let α be any assignment to the variables $x_1, \ldots, x_m$, and let w be any bottom world of a group γ that represents α. Then, by construction of $[\psi]$, we know that ψ is true under α if and only if $[\psi]$ is true in w.

Assume that the claim holds for $i = j+1$. We show that then the claim also holds for $i = j$. Let α be any assignment to the variables $x_1, \ldots, x_{j-1}$ and let w be a bottom world of a group that agrees with α. We show that the formula $Q_j \ldots Q_m . \psi$ is true under α if and only if $[Q_j] \ldots [Q_m][\psi]$ is true in w.

First, assume that $Q_j \ldots Q_m . \psi$ is true under α. Consider the case where $Q_j = \forall$. Then for both assignments $\alpha' \supseteq \alpha$ to the variables $x_1, \ldots, x_j$, formula $Q_{j+1} \ldots Q_m . \psi$ is true under α'. Now, by assumption, we know that for any bottom world w' of a group that agrees with α – so in particular for all bottom worlds w' that are R_j-reachable from w – formula $[Q_{j+1}] \ldots [Q_m][\psi]$ is true in w'. Since $[Q_j] = B_j$, this means that $[Q_j] \ldots [Q_m][\psi]$ is true in w. The case where $Q_j = \exists$ is analogous.

Next, assume that $Q_j \ldots Q_m . \psi$ is not true under α. Consider the case where $Q_j = \forall$. Then there is some assignment $\alpha' \supseteq \alpha$ to the variables $x_1, \ldots, x_j$, such that $Q_{j+1} \ldots Q_m . \psi$ is not true under α'. Now, by assumption, we know that for any bottom world w' of a group that agrees with α – so in particular for some bottom world w' that is R_j-reachable from w – formula $[Q_{j+1}] \ldots [Q_m][\psi]$ is not true in w'. Since $[Q_j] = B_j$, this means that $[Q_j] \ldots [Q_m][\psi]$ is not true in w. The case where $Q_j = \exists$ is analogous.

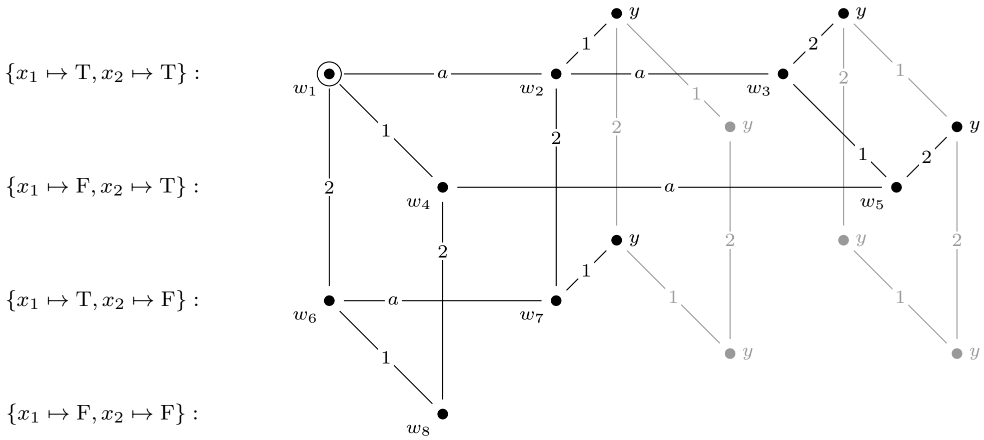
Hence, the claim holds for the case that $i = j$. Now, by induction, the claim holds for the case that $i = 1$, and hence it follows that $\varphi \in$ TQBF if and only if $R(\varphi) \in$ DBU. Since this reduction runs in polynomial time, we can conclude that DBU is PSPACE-hard. □

THEOREM 2. *DBU is* PSPACE*-complete.*

PROOF. In order to show PSPACE-membership for the problem DBU, we can modify the polynomial-space algorithm given by Aucher and Schwarzentruber [3]. Their algorithm works for the problem of checking whether a given (single-pointed) epistemic model makes a given DEL-formula true, where the formula contains event models that can be multi-pointed, but that have no postconditions. In order to make the algorithm work for multi-pointed epistemic models, we can simply call the algorithm several times, once for each of the designated worlds. Also, a modification is needed to deal with postconditions. The algorithm checks the truth of a formula by inductively calling itself for subformulas. In order to deal with postconditions, only the case where the formula is a propositional variable needs to be modified. This modification is rather straightforward. For more details, we refer to [37]. □

4.2 Parameterized Complexity Results

Next, we provide a parameterized complexity analysis of DBU.



$\{x_1 \mapsto \mathrm{T}, x_2 \mapsto \mathrm{T}\}$:
$\{x_1 \mapsto \mathrm{F}, x_2 \mapsto \mathrm{T}\}$:
$\{x_1 \mapsto \mathrm{T}, x_2 \mapsto \mathrm{F}\}$:
$\{x_1 \mapsto \mathrm{F}, x_2 \mapsto \mathrm{F}\}$:
w_1
w_2
w_3
w_4
w_5
w_6
w_7
w_8
a
y
1
2

Figure 1: Example for the reduction in the proof of Theorem 1; a final updated model for a quantified Boolean formula with variables x_1 and x_2.

4.2.1 Parameters for DBU

We consider the following parameters for DBU. For each subset $\kappa \subseteq \{a, c, e, f, o, p, u\}$ we consider the parameterized variant κ-DBU of DBU, where the parameter is the sum of the values for the elements of κ as specified in Table 1. For instance, the problem $\{a\}$-DBU is parameterized by the number of agents. Even though technically speaking there is only one parameter, we will refer to each of the elements of κ as parameters.

For the modal depth of a formula we count the maximum number of nested occurrences of operators B_a. Formally, we define the modal depth $d(\varphi)$ of a formula φ (in $\mathcal{L}_B$) recursively as follows.

$$d(\varphi) = \begin{cases} 0 & \text{if } \varphi = p \in P \text{ is a proposition;} \\ \max\{d(\varphi_1), d(\varphi_2)\} & \text{if } \varphi = \varphi_1 \wedge \varphi_2; \\ d(\varphi_1) & \text{if } \varphi = \neg\varphi_1; \\ 1 + d(\varphi_1) & \text{if } \varphi = B_a\varphi_1. \end{cases}$$

For the size of a formula we count the number of occurrences of propositions and logical connectives. Formally, we define the size $s(\varphi)$ of a formula φ (in $\mathcal{L}_B$) recursively as follows.

$$s(\varphi) = \begin{cases} 1 & \text{if } \varphi = p \in P \text{ is a proposition;} \\ 1 + s(\varphi_1) + s(\varphi_2) & \text{if } \varphi = \varphi_1 \wedge \varphi_2; \\ 1 + s(\varphi_1) & \text{if } \varphi = \neg\varphi_1; \\ 1 + s(\varphi_1) & \text{if } \varphi = B_a\varphi_1. \end{cases}$$

4.2.2 Intractability Results

In the following, we show fixed-parameter intractability for several parameterized versions of DBU. We will mainly use the parameterized complexity classes W[1] and para-NP to show intractability, i.e., we will show hardness for these classes. Note that we could additionally use the class para-PSPACE [21] to give stronger intractability results. For instance, the proof of Theorem 1 already shows that $\{p\}$-DBU is para-PSPACE hard, since the reduction in this proof uses a constant number of propositions. However, since in this paper we are mainly interested in the border between fixed-parameter tractability and intractability, we will not focus on the subtle differences in the degree of intractability, and restrict ourselves to showing W[1]-hardness and para-NP-hardness. This is also the reason why we will not show membership for any of the (parameterized) intractability classes; showing hardness suffices to indicate intractability. For the following proofs we use the well-known satisfiability problem SAT for propositional formulas. The problem SAT is NP-complete [14, 30]. Moreover, hardness for SAT holds even when restricted to propositional formulas that are in 3CNF.

Param.	**Description**
a	number of agents
c	maximum size of the preconditions
e	maximum number of events in the event models
f	size of the formula
o	modal depth of the formula, i.e., the order parameter
p	number of propositions in P
u	number of actions, i.e., the number of updates

Table 1: Overview of the different parameters for DBU**.**

PROPOSITION 3. $\{a, c, e, f, o\}$-DBU *is* para-NP*-hard.*

PROOF. To show para-NP-hardness, we specify a polynomial-time reduction R from SAT to DBU, where parameters a, c, e, f, and o have constant values. Let φ be a propositional formula with $var(\varphi) = \{x_1, \dots, x_m\}$. Without loss of generality we assume that φ is a 3CNF formula with clauses c_1 to c_l.

The general idea behind this reduction is that we use the worlds in the final updated model (that results from updating the initial state with the actions – which are specified by the reduction) to list all possible assignments to $var(\varphi)$, by setting the propositions (corresponding to the variables in $var(\varphi)$) to true and false accordingly. Then checking whether formula φ is satisfiable can be done by checking whether φ is true in any of the worlds. Actions a_1 to a_m are used to create a corresponding world for each possible assignment to $var(\varphi)$. Furthermore, to keep the formula that we check in the final updated model of constant size, we sequentially check the truth of each clause c_i and encode whether the clauses are true with an additional variable x_{m+1}. This is done by actions a_{m+1} to a_{m+l}. In the final updated model, variable x_{m+1} will only be true in a world, if it makes clauses c_1 to c_l true, i.e., if it makes formula φ true.

For more details, we refer to [37]. □

Proposition 4. $\{c, e, f, o, p\}$-DBU *is* para-NP*-hard.*

Proof. To show para-NP-hardness, we specify a polynomial-time reduction R from Sat to DBU, where parameters c, e, f, o, and p have constant values. Let φ be a propositional formula with $var(\varphi) = \{x_1, \dots, x_m\}$. The general idea behind this reduction is similar to the reduction in the proof of Theorem 1. Again we use groups of worlds to represent particular assignments to the variables in φ. Here, there is only relation R_b between the different groups. Furthermore, to keep the formula that we check in the final updated model of constant size, we sequentially check the truth of each clause c_i and encode whether the clauses are true with an additional variable z. This is done by actions a_{m+1} to a_{m+l}. Action a_{m+j} (corresponding to clause j) marks each group of worlds (which represents a particular assignment to the variables in φ) that 'satisfies' clauses 1 to j. (This marking happens by means of an R_c-accessible world where z is true.) Then, in the final updated model, there will only be such a marked group if all clauses, and hence the whole formula, is satisfiable.

For more details, we refer to [37]. □

Proposition 5. $\{a, e, f, o, p\}$-DBU *is* para-NP*-hard.*

Proof. To show para-NP-hardness, we specify a polynomial-time reduction R from Sat to DBU, where parameters a, e, f, o and p have constant values. Let φ be a propositional formula with $var(\varphi) = \{x_1, \dots, x_m\}$. The reduction is based on the same principle as the one used in the proof of Proposition 4. To keep the number of agents constant, we use a different construction to represent the variables in $var(\varphi)$. We encode the variables by a string of worlds that are connected by alternating relations R_a and R_b.

Furthermore, we keep the size of the formula (and consequently the modal depth of the formula) constant by encoding the satisfiability of the formula with a single proposition. We do this by adding an extra action a_{m+1}. Action a_{m+1} makes sure that each group of worlds that represents a satisfying assignment for the given formula, will have an R_c relation from a world that is R_b-reachable from the designated world to a world where proposition z^* is true.

For more details, we refer to [37]. □

We consider the following parameterized problem, that we will use in our proof of Proposition 6. This problem is W[1]-complete [19].

$\{k\}$-Multicolored Clique
Instance: A graph G, and a vertex-coloring $c : V(G) \to \{1, 2, \dots, k\}$ for G.
Parameter: k.
Question: Does G have a clique of size k including vertices of all k colors? That is, are there $v_1, \dots, v_k \in V(G)$ such that for all $1 \leq i < j \leq k : \{v_i, v_j\} \in E(G)$ and $c(v_i) \neq c(v_j)$?

Proposition 6. $\{a, c, f, o, u\}$-DBU *is* W[1]*-hard.*

Proof. We specify an fpt-reduction R from $\{k\}$-Multicolored Clique to $\{a, c, f, o, u\}$-DBU. Let (G, c) be an instance of $\{k\}$-Multicolored Clique, where $G = (N, E)$. The general idea behind this reduction is that we use the worlds in the model to list all k-sized subsets of the vertices in the graph with k different colors, where each individual world represents a particular k-subset of vertices in the graph (with k different colors). Then we encode (in the model) the existing edges between these nodes (with particular color endings), and in the final updated model we check whether there is a world corresponding to a k-subset of vertices that is pairwise fully connected with edges. This is only the case when G has a k-clique with k different colors.

For more details, we refer to [37]. □

Proposition 7. $\{c, o, p, u\}$-DBU *is* W[1]*-hard.*

Proof. We specify the following fpt-reduction R from $\{k\}$-WSat[2CNF] to $\{c, o, p, u\}$-DBU. We sketch the general idea behind the reduction. Let φ be a propositional formula with $var(\varphi) = \{x_1, \dots, x_m\}$. Then let φ' be the formula obtained from φ, by replacing each occurrence of x_i with $\neg x_i$. We note that φ is satisfiable by some assignment α that sets k variables to true if and only if φ' is satisfiable by some assignment α' that sets $m - k$ variables to true, i.e., that sets k variables to false. We use the reduction to list all possible assignments to $var(\varphi') = var(\varphi)$ that set $m - k$ variables to true. We represent each possible assignment to $var(\varphi)$ that sets $m - k$ variables to true as a group of worlds, like in the proof of Theorem 1. (In fact, due to the details of the reduction, in the final updated model, there will be several identical groups of worlds for each of these assignments).

For more details, we refer to [37]. □

Proposition 8. $\{a, f, o, p, u\}$-DBU *is* W[1]*-hard.*

Proof. We specify the following fpt-reduction R from $\{k\}$-WSat[2CNF] to $\{a, f, o, p, u\}$-DBU. We modify the reduction in the proof of Proposition 7 to keep the values of parameters a and f constant. After these modifications, the value of parameter c will no longer be constant. To keep the number of agents constant, we use the same strategy as in the reduction in the proof of Proposition 5, where variables $x_i, \dots, x_m$ are represented by strings of worlds with alternating relations R_b and R_a. Just like in the proof of Proposition 5, the size of the formula (and consequently the modal depth of the formula) is kept constant by encoding the satisfiability of the formula with a single proposition. Then each group of worlds that represents a satisfying assignment for the given formula, will have an R_c relation from a world that is R_b-reachable from the designated world to a world where proposition z^* is true.

For more details, we refer to [37]. □

4.2.3 Tractability Results

Next, we turn to a case that is fixed-parameter tractable.

THEOREM 9. $\{e, u\}$-DBU *is fixed-parameter tractable.*

PROOF. We present the following fpt-algorithm that runs in time $e^u \cdot p(|x|)$, for some polynomial p, where e is the maximum number of events in the actions and u is the number of updates, i.e., the number of actions.

As a subroutine, the algorithm checks whether a given basic epistemic formula φ holds in a given epistemic model M, i.e., whether $M \models \varphi$. It is well-known that model checking for basic epistemic logic can be done in time polynomial in the of M plus the size of φ (see e.g. [7]).

Let $x = (P, \mathcal{A}, i, s_0, a_1, \ldots, a_f, \varphi)$ be an instance of DBU. First the algorithm computes the final updated model $s_f = s_0 \otimes a_1 \otimes \cdots \otimes a_f$ by sequentially performing the updates. For each i, s_i is defined as $s_{i-1} \otimes a_i$. The size of each s_i is upper bounded by $O(|s_0| \cdot e^u)$, so for each update checking the preconditions can be done in time polynomial in $e^u \cdot |x|$. This means that computing s_f can be done in fpt-time.

Then, the algorithm decides whether φ is true in s_f. This can be done in time polynomial in the size of s_f plus the size of φ. We know that $|s_f| + |\varphi|$ is upper bounded by $O(|s_0| \cdot e^u) + |\varphi|$, thus upper bounded by $e^u \cdot p(|x|)$, for some polynomial p. Therefore, deciding whether φ is true in s_f is fixed-parameter tractable. Hence, the algorithm decides whether $x \in$ DBU and runs in fpt-time. □

4.2.4 Overview of the Results

We showed that DBU is PSPACE-complete, we presented several parameterized intractability results (W[1]-hardness and para-NP-hardness) and we presented one fixed-parameter tractable result, namely for $\{e, u\}$-DBU. In Figure 2, we present a graphical overview of our results and the consequent border between fpt-tractability and fpt-intractability for the problem DBU. We leave $\{a, c, p\}$-DBU and $\{c, f, p, u\}$-DBU as open problems for future research.

5. DISCUSSION & CONCLUSIONS

We presented the DYNAMIC BELIEF UPDATE model as a computational-level model of ToM and analyzed its complexity. The aim of our model was to provide a formal approach that can be used to interprete and evaluate the meaning and veridicality of various complexity claims in the cognitive science and philosophy literature concerning ToM. In this way, we hope to contribute to disentangling debates in cognitive science and philosophy regarding the complexity of ToM.

In Section 4.1, we proved that DBU is PSPACE-complete. This means that (without additional constraints), there is no algorithm that computes DBU in a reasonable amount of time. In other words, without restrictions on its input domain, the model is computationally too hard to serve as a plausible explanation for human cognition. This may not be surprising, but it is a first formal proof backing up this claim, whereas so far claims of intractability in the literature remained informal.

Informal claims about what constitutes sources of intractability abound in cognitive science. For instance, it seems to be folklore that the 'order' of ToM reasoning (i.e., that I think that you think that I think . . .) is a potential source of intractability. The fact that people have difficulty understanding higher-order theory of mind [20, 29, 32, 44] is not explained by the complexity results for parameter o – the modal depth of the formula that is being considered, in other words, the order parameter. Already for a formula with modal depth one, DBU is NP-hard; so $\{o\}$-DBU is not fixed-parameter tractable. On the basis of our results we can only conclude that DBU is fixed-parameter tractable for the order parameter in combination with parameters e and u. But since DBU is fp-tractable for the smaller parameter set $\{e, u\}$, this does not indicate that the order parameter is a source of complexity. This does not mean it may not be a source of difficulty for human ToM performance. After all, tractable problems can be too resource-demanding for humans for other reasons than computational complexity (e.g., due to stringent working-memory limitations).

Surprisingly, we only found one (parameterized) tractability result for DBU. We proved that for parameter set $\{e, u\}$ – the maximum number of events in an event model and the number of updates, i.e., the number of event models – DBU is fixed-parameter tractable. Given a certain instance x of DBU, the values of parameters e and u (together with the size of initial state s_0) determine the size of the final updated model (that results from applying the event models to the initial state). Small values of e and u thus make sure that the final updated model does not blow up too much in relation to the size of the initial model. The result that $\{e, u\}$-DBU is fp-tractable indicates that the size of the final updated model can be a source of intractability (cf. [39, 40]).

The question arises how we can interpret parameters e and u in terms of their cognitive counterparts. To what aspect of ToM do they correspond, and moreover, can we assume that they have small values in (many) real-life situations? If this is indeed the case, then restricting the input domain of the model to those inputs that have sufficiently small values for parameters e and u will render our model tractable, and we can then argue that (at least in terms of its computational complexity) it is a cognitively plausible model.

In his formalizations of the false belief task Bolander [8] indeed used a limited amount of actions with a limited amount of events in each action (he used a maximum of 4). This could, however, be a consequence of the over-simplification (of real-life situations) used in experimental tasks. Whether these parameters in fact have sufficiently small values in real life, is an empirical hypothesis that can (in principle) be tested experimentally. However, it is not straightforward how to interpret these formal aspects of the model in terms of their cognitive counterparts. The associations that the words *event* and *action* trigger with how we often use these words in daily life, might adequately apply to some degree, but could also be misleading. A structural way of interpreting these parameters is called for. We think this is an interesting topic for future research.

Besides the role that our results play in the investigation of (the complexity) of ToM our results are also of interest in and of themselves. The results in Theorems 1 and 2 resolve an open question in the literature about the computational complexity of DEL. Aucher and Schwarzentruber [3] already showed that the model checking problem for DEL, in general, is PSPACE-complete. However, their proof for

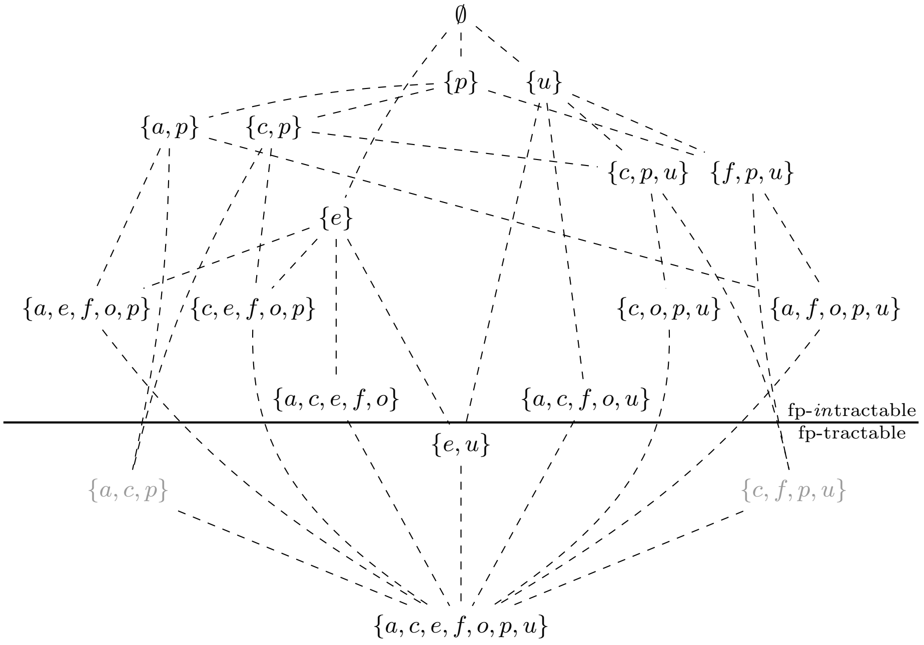


Figure 2: Overview of the parameterized complexity results for the different parameterizations of DBU, and the line between fp-tractability and fp-intractability (under the assumption that the cases for $\{a,c,p\}$ and $\{c,f,p,u\}$ are fp-tractable).

PSPACE-hardness does not work when the input domain is restricted to S5 (or KD45) models and their hardness proof also relies on the use of multi-pointed models (which in their notation is captured by means of a union operator). With our proof of Theorem 1, we show that DEL model checking is PSPACE-hard even when restricted to single-pointed S5 models. Furthermore, the novelty of our aproach lies in the fact that we apply parameterized complexity analysis to dynamic epistemic logic, which is still a rather unexplored area.

Acknowledgements

We thank the reviewers for their comments. We thank Thomas Bolander, Nina Gierasimczuk, Ronald de Haan and Martin Holm Jensen and the members of the Computational Cognitive Science group at the Donders Centre for Cognition for discussions and feedback.

6. REFERENCES

[1] I. Apperly. *Mindreaders: the cognitive basis of "Theory of Mind"*. Psychology Press, 2011.

[2] G. Aucher. An internal version of epistemic logic. *Studia Logica*, 94(1):1–22, 2010.

[3] G. Aucher and F. Schwarzentruber. On the complexity of dynamic epistemic logic. In *Proceedings of the Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, 2013.

[4] A. Baltag, L. S. Moss, and S. Solecki. The logic of public announcements, common knowledge, and private suspicions. In *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, 1998.

[5] S. Baron-Cohen, A. M. Leslie, and U. Frith. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46, 1985.

[6] J. van Benthem. *Logical Dynamics of Information and Interaction.* Cambridge University Press, Cambridge, 2011.

[7] P. Blackburn, J. van Benthem, et al. Modal logic: A semantic perspective. *Handbook of modal logic*, 3:1–84, 2006.

[8] T. Bolander. Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In *Proceedings of European Conference on Social Intelligence (ECSI 2014)*, pages 87–107, 2014.

[9] T. Bolander and M. B. Andersen. Epistemic planning for single and multi-agent systems. *Journal of Applied Non-Classical Logics*, 21(1):9–34, 2011.

[10] T. Bolander, M. H. Jensen, and F. Schwarzentruber. Complexity results in epistemic planning. In *Proceedings of 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 2015.

[11] T. Braüner. Hybrid-logical reasoning in false-belief tasks. In B. Schipper, editor, *Proceedings of the Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, 2013.

[12] N. Chater, J. B. Tenenbaum, and A. Yuille. Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7):287–291, 2006.

[13] C. Cherniak. *Minimal rationality.* MIT Press, 1990.

[14] S. A. Cook. The complexity of theorem proving procedures. In *Proceedings of the 3rd Annual ACM Symposium on the Theory of Computing (STOC)*,

pages 151–158. ACM, 1971.

[15] C. Dégremont, L. Kurzen, and J. Szymanik. Exploring the tractability border in epistemic tasks. *Synthese*, 191(3):371–408, 2014.

[16] H. van Ditmarsch, W. van der Hoek, and B. P. Kooi. *Dynamic Epistemic Logic*. Springer, 2008.

[17] R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Monographs in Computer Science. Springer, New York, 1999.

[18] R. G. Downey and M. R. Fellows. *Fundamentals of Parameterized Complexity*. Texts in Computer Science. Springer, 2013.

[19] M. R. Fellows, D. Hermelin, F. A. Rosamond, and S. Vialette. On the parameterized complexity of multiple-interval graph problems. *Theoretical Computer Science*, 410(1):53–61, 2009.

[20] L. Flobbe, R. Verbrugge, P. Hendriks, and I. Krämer. Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, 17(4):417–442, 2008.

[21] J. Flum and M. Grohe. Describing parameterized complexity classes. *Information and Computation*, 187(2):291–319, 2003.

[22] J. Flum and M. Grohe. *Parameterized Complexity Theory*, volume XIV of *Texts in Theoretical Computer Science. An EATCS Series*. Springer, Berlin, 2006.

[23] U. Frith. Mind blindness and the brain in autism. *Neuron*, 32(6):969–979, 2001.

[24] M. Frixione. Tractable competence. *Minds and Machines*, 11(3):379–397, 2001.

[25] N. Gierasimczuk and J. Szymanik. A note on a generalization of the muddy children puzzle. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, 2011.

[26] G. Gigerenzer. Why heuristics work. *Perspectives on psychological science*, 3(1):20–29, 2008.

[27] W. F. G. Haselager. *Cognitive Science and Folk Psychology: The Right Frame of Mind*. Sage Publications, 1997.

[28] A. M. Isaac, J. Szymanik, and R. Verbrugge. Logic and complexity in cognitive science. In A. Baltag and S. Smets, editors, *Johan van Benthem on Logic and Information Dynamics*, volume 5 of *Outstanding Contributions to Logic*, pages 787–824. Springer International Publishing, 2014.

[29] P. Kinderman, R. Dunbar, and R. P. Bentall. Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, 89(2):191–204, 1998.

[30] L. A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9(3):265–266, 1973.

[31] S. C. Levinson. On the human 'interaction engine'. In N. J. Enfield and S. C. Levinson, editors, *Roots of human sociality: Culture, cognition and interaction*, pages 39–69. Oxford: Berg, 2006.

[32] M. Lyons, T. Caldwell, and S. Shultz. Mind-reading and manipulation – is machiavellianism related to theory of mind? *Journal of Evolutionary Psychology*, 8(3):261–274, 2010.

[33] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: WH Freeman, 1982.

[34] S. Nichols and S. P. Stich. *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Clarendon Press/Oxford University Press, 2003.

[35] R. Niedermeier. *Invitation to Fixed-Parameter Algorithms*. Oxford Lecture Series in Mathematics and its Applications. Oxford University Press, 2006.

[36] A. Perea. *Epistemic Game Theory: reasoning and choice*. Cambridge University Press, 2012.

[37] I. van de Pol. How Difficult is it to Think that you Think that I Think that ...? *A DEL-based Computational-level Model of Theory of Mind and its Complexity*. Master's thesis, University of Amsterdam, the Netherlands, 2015.

[38] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(04):515–526, 1978.

[39] I. van Rooij, P. Evans, M. Müller, J. Gedge, and T. Wareham. Identifying sources of intractability in cognitive models: An illustration using analogical structure mapping. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 915–920, 2008.

[40] I. van Rooij and T. Wareham. Parameterized complexity in cognitive modeling: Foundations, applications and opportunities. *The Computer Journal*, 51(3):385–404, 2008.

[41] I. van Rooij, C. D. Wright, J. Kwisthout, and T. Wareham. Rational analysis, intractability, and the prospects of 'as if'-explanations. *Synthese*, pages 1–20, 2014.

[42] I. van Rooij. The tractable cognition thesis. *Cognitive Science*, 32(6):939–984, 2008.

[43] I. van Rooij, C. D. Wright, and T. Wareham. Intractability and the use of heuristics in psychological explanations. *Synthese*, 187(2):471–487, 2012.

[44] J. Stiller and R. I. Dunbar. Perspective-taking and memory capacity predict social network size. *Social Networks*, 29(1):93–104, 2007.

[45] L. J. Stockmeyer and A. R. Meyer. Word problems requiring exponential time (preliminary report). In *Proceedings of the 5th Annual ACM Symposium on the Theory of Computing (STOC)*, pages 1–9. ACM, 1973.

[46] J. K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13(03):423–445, 1990.

[47] R. Verbrugge. Logic and social cognition: the facts matter, and so do computational models. *Journal of Philosophical Logic*, 38(6):649–680, 2009.

[48] H. M. Wellman, D. Cross, and J. Watson. Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3):655–684, 2001.

[49] H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.

[50] T. W. Zawidzki. *Mindshaping: A New framework for understanding human social cognition*. MIT Press, 2013.

A Dynamic Epistemic Framework for Conformant Planning

Quan Yu[1,2], Yanjun Li[3,4] and Yanjing Wang[*3]

[1]Department of Computer Science, Sun Yat-sen University, China

[2]Qiannan Normal College for Nationalities, China

[3]Department of Philosophy, Peking University, China

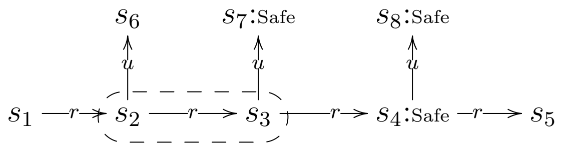
[4]Faculty of Philosophy, University of Groningen, The Netherlands

ABSTRACT

In this paper, we introduce a lightweight dynamic epistemic logical framework for automated planning under initial uncertainty. We reduce plan verification and conformant planning to model checking problems of our logic. We show that the model checking problem of the iteration-free fragment is PSPACE-complete. By using two non-standard (but equivalent) semantics, we give novel model checking algorithms to the full language and the iteration-free language.

1. INTRODUCTION

Conformant planning is the problem of finding a linear plan (a sequence of action) to achieve a goal in presence of uncertainty about the initial state (cf. [28]). For example, suppose that you are a rookie spy trapped in a foreign hotel with the following map at hand:[1]



Now somebody spots you and sets up the alarm. In this case you need to move fast to one of the safe hiding places marked in the map (i.e., s_7, s_8 and s_4). However, since you were in panic, you lost your way and you are not sure whether you are at s_2 or s_3 (denoted by the circle in the above graph). Now what should you do in order to reach a safe place quickly? Clearly, merely moving r or moving u may not guarantee your safety given the uncertainty. A simple plan is to move r first and then u, since this plan will take you to a safe place, no matter where you actually are initially. This plan is *conformant* since it does not require any feedback during the execution and it should work in presence of uncertainty about the initial state. More generally, a conformant plan should also work given actions with non-deterministic effects. Such a conformant plan is crucial when there are no feedbacks/observations available during the execution of the plan.[2] Note that since no information is provided during the execution, the conformant plan is simply a finite sequence of actions without any conditional moves.

As discussed in [9, 24], conformant planning can be reduced to classical planning, the planning problem without any initial uncertainty, over the space of *belief states*. Intuitively, a belief state is a subset of the state space, which records the uncertainty during the execution of a plan, e.g., $\{s_2, s_3\}$ is an initial belief state in the above example. In order to make sure a goal is achieved eventually, it is crucial to track the transitions of belief states during the execution of the plan, and this may traverse exponentially many belief states in the size of the original state space. As one may expect, conformant planning is computationally harder than classical planning. The complexity of checking the existence of a conformant plan is EXPSPACE-complete in the size of the variables generating the state space [18]. In the literature, people proposed compact and implicit representations of the belief spaces, such as OBDD [13, 15, 14] and CNF [30], and different heuristics are used to guide the search for a plan, e.g., [11, 12].

Besides the traditional AI approaches, we can also take an epistemic-logical perspective on planning in presence of initial uncertainties, based on dynamic epistemic logic (DEL) (cf. e.g., [31]). The central philosophy of DEL takes the meaning of an action as the change it brings to the knowledge of the agents. Intuitively, this is what we need to track the belief states during the execution of a plan[3]. Indeed, in recent years, there has been a growing interest in using DEL to handle multi-agent planning with knowledge goals (cf. e.g., [7, 23, 1, 3, 34, 25]), while the traditional AI planning focuses on the single-agent case. In particular, the event models of DEL (cf. [6]) are used to handle non-public actions that may cause different knowledge updates to different agents. In these DEL-based planning frameworks, states are epistemic models, actions are event models and the state transitions are implicitly encoded by the update product which computes a new epistemic model based on an epistemic model and an event model.

One advantage of this approach is its expressiveness in handling scenarios which require reasoning about agents' higher-order knowledge about each other in presence of partially observable actions. However, this expressiveness comes at a price, as shown in [7, 4], that multi-agent epistemic plan-

[*]Corresponding author

[1] It is a variant of the running example in [33].

[2] In many other cases, feedbacks may be just too 'expensive' to obtain during a plan aiming for quick actions [8].

[3] Here the belief states are actually about knowledge in epistemic logic.

ning is undecidable in general. Many interesting decidable fragments are found in the literature [7, 23, 34, 2], which suggests that the single-agent cases and restrictions on the form of event models are the key to decidability. However, if we focus on the single-agent planning, a natural question arises: how do we compare such DEL approaches with the traditional AI planning? It seems that the DEL-based approaches are more suitable for planning with actions that change (higher-order) knowledge rather than planning with fact-changing actions, although the latter type of actions can also be handled in DEL. Moreover, the standard models of DEL are purely epistemic thus do not encode the temporal information of available actions directly. This may limit the applicability of such approaches to planning problems based on transition systems.

In this paper, we tackle the standard single-agent conformant planning problem over transition systems, by using the core idea of DEL, but not its standard formalism. Our formal framework is based on the logic proposed by Wang and Li in [33], where the model is simply a transition system with initial uncertainty as in the motivating example, and an action is interpreted in the semantics as an update on the uncertainty of the agent. Our contributions are summarized as follows:

- A lightweight dynamic epistemic framework with a simple language and a complete axiomatization.
- Non-trivial reduction of conformant planning to a model checking problem using our language with programs.
- Two novel model checking algorithms based on two alternative semantics for the proposed logic, which make the context-dependency in the original semantics explicit.
- The complexity of model checking the iteration-free fragment of our language is PSPACE-complete. The model checking problem of the full language is in EXPTIME. The model checking problem of the conformant planning is in PSPACE.

The last result may sound contradictory to the aforementioned result that the complexity of conformant planning is EXPSPACE-complete. Actually, the apparent contradiction is due to the fact that the EXPSPACE complexity result is based on the number of *state variables* which require an exponential blow up to generate an explicit transition system that we use here. We will come back to this issue at the end of Section 4.3.

Our approach has the following advantages compared to the existing planning approaches:

- The planning goals can be specified as arbitrary formulas in an epistemic language. Extra plan constraints (e.g., what actions to use) can be expressed explicitly by programs in the language. Therefore it may cover a richer class of (conformant) planning problems compared to the traditional AI approach where a goal is Boolean.[4]
- The plans can be specified as regular expressions with tests in terms of arbitrary EPDL formulas, which generalizes the knowledge-based programs in [17, 21].
- By reducing conformant planning to a model checking problem in an explicit logical language, we also see the subtleties hidden in the planning problem. In principle, there are various model checking techniques to be applied to conformant planning based on this reduction.
- Our logical language and models are very simple compared to the standard action-model based DEL approach, yet we can encode the externally given executability of the actions in the model, inspired by epistemic temporal logic (ETL) [16, 26].
- Our approach is flexible enough to provide, in the future, a unified platform to compare different planning problems under uncertainty. By studying different fragments of the logical language and model classes, we may categorize planning problems according to their complexity.

The rest of the paper is organized as follows: We introduce our basic logical framework and its axiomatization in Section 2, and extend it in Section 3 with programs to handle the conformant planning. The complexity analysis of the model checking problems is in Section 4 and we conclude in Section 5 with future directions.

2. BASIC FRAMEWORK

2.1 Epistemic action language

To talk about the knowledge of the agent during an execution of a plan, we use the following language proposed in [33].

DEFINITION 2.1 (EPISTEMIC ACTION LANGUAGE (EAL)). *Given a countable set* A *of action symbols and a countable set* P *of atomic proposition letters , the language* $\mathtt{EAL}^{\mathtt{A}}_{\mathtt{P}}$ *is defined as follows:*[5]

$$\phi ::= \top \mid p \mid \neg\phi \mid (\phi \wedge \phi) \mid [a]\phi \mid K\phi,$$

where $p \in \mathtt{P}$, $a \in \mathtt{A}$. *The following standard abbreviations are used:* $\bot := \neg\top$, $\phi \vee \psi := \neg(\neg\phi \wedge \neg\psi)$, $\phi \to \psi := \neg\phi \vee \psi$, $\langle a\rangle\phi := \neg[a]\neg\phi$, $\hat{K}\phi := \neg K\neg\phi$.

$K\phi$ says that the agent knows that ϕ, and $[a]\phi$ expresses that if the agent can move forward by action a, then after doing a, ϕ holds. Throughout the paper, we fix some P and A, and refer to $\mathtt{EAL}^{\mathtt{A}}_{\mathtt{P}}$ by EAL.

The size of EAL-formulas (notation $|\varphi|$) is defined inductively: $|\top| = |p| = 1$; $|\neg\phi| = 1 + |\phi|$; $|\phi \wedge \psi| = 1 + |\phi| + |\psi|$; $|K\phi| = |[a]\phi| = 1 + |\phi|$. The set of subformulas of $\phi \in \mathtt{EAL}$, denoted as $sub(\phi)$, is defined as usual.

DEFINITION 2.2 (UNCERTAINTY MAP). *Given* P *and* A, *a (multimodal) Kripke model* $\mathcal{N}$ *is a tuple* $\langle \mathcal{S}, \{\mathcal{R}_a \mid a \in \mathtt{A}\}, \mathcal{V}\rangle$, *where* $\mathcal{S}$ *is a non-empty set of states,* $\mathcal{R}_a \subseteq \mathcal{S} \times \mathcal{S}$ *is a binary relation labelled by* a, $\mathcal{V} : \mathcal{S} \to 2^{\mathtt{P}}$ *is a valuation function. An uncertainty map* $\mathcal{M}$ *is a Kripke model* $\langle \mathcal{S}, \{\mathcal{R}_a \mid a \in \mathtt{A}\}, \mathcal{V}\rangle$ *with a non-empty set* $\mathcal{U} \subseteq \mathcal{S}$. *Given an uncertainty map* $\mathcal{M}$, *we refer to its components by* $\mathcal{S}_{\mathcal{M}}$, $\mathcal{R}_{a\mathcal{M}}$, $\mathcal{V}_{\mathcal{M}}$, *and* $\mathcal{U}_{\mathcal{M}}$. *A pointed uncertainty map* $\mathcal{M}, s$ *is an uncertainty map* $\mathcal{M}$ *with a designated state* $s \in \mathcal{U}_{\mathcal{M}}$. *We write* $s \xrightarrow{a} t$ *for* $(s, t) \in \mathcal{R}_a$.

[4] The goal in the standard conformant planning is simply a set of different valuations of basic propositional variables. Our approach can even handle epistemic goals in negative forms, e.g., we want to make sure the agent knows something but does not know too much in the end.

[5] We do need unboundedly many action symbols to encode the desired problem in the later discussion of model checking complexity.

Intuitively, a Kripke model encodes a map (transition system) and the uncertainty set $\mathcal{U}$ encodes the uncertainty that the agent has about where he is in the map. The graph mentioned at the beginning of the introduction is a typical example of an uncertainty map. Note that there may be non-deterministic transitions in the model, i.e., there may be $t_1 \neq t_2$ such that $s \xrightarrow{a} t_1$ and $s \xrightarrow{a} t_2$ for some s, t_1, t_2.

REMARK 1. *It is crucial to notice that the designated state in a pointed uncertainty map must be one of the states in the uncertainty set.*

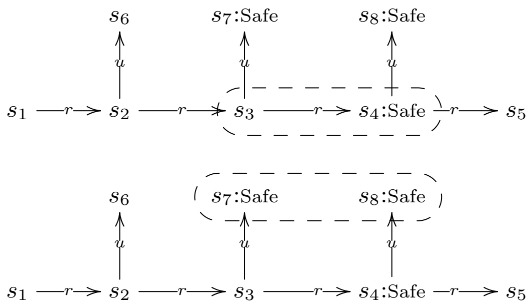
DEFINITION 2.3 (SEMANTICS). *Given any uncertainty map $\mathcal{M} = \langle \mathcal{S}, \{\mathcal{R}_a \mid a \in \mathtt{A}\}, \mathcal{V}, \mathcal{U}\rangle$ and any state $s \in \mathcal{U}$, the semantics is defined as follows:*

$$\boxed{\begin{array}{lll}
\mathcal{M}, s \vDash \top & & \textit{always} \\
\mathcal{M}, s \vDash p & \iff & s \in \mathcal{V}(p) \\
\mathcal{M}, s \vDash \neg\phi & \iff & \mathcal{M}, s \nvDash \phi \\
\mathcal{M}, s \vDash \phi \wedge \psi & \iff & \mathcal{M}, s \vDash \phi \textit{ and } \mathcal{M}, s \vDash \psi \\
\mathcal{M}, s \vDash [a]\phi & \iff & \forall t \in S : s \xrightarrow{a} t \textit{ implies } \mathcal{M}|^a, t \vDash \phi \\
\mathcal{M}, s \vDash K\phi & \iff & \forall u \in \mathcal{U} : \mathcal{M}, u \vDash \phi
\end{array}}$$

where $\mathcal{M}|^a = \langle \mathcal{S}, \{\mathcal{R}_a \mid a \in \mathtt{A}\}, \mathcal{V}, \mathcal{U}|^a\rangle$ and $\mathcal{U}|^a = \{r' \mid \exists r \in \mathcal{U}$ such that $r \xrightarrow{a} r'\}$. We say ϕ is valid (notation: $\vDash \phi$) if it is true on all the pointed uncertainty maps. For a action sequence $\sigma = a_1 \ldots a_n$, we write $\mathcal{U}|^\sigma$ for $(\ldots((\mathcal{U}|^{a_1})|^{a_2})\ldots)|^{a_n}$. and write $\mathcal{M}|^\sigma$ for $(\ldots((\mathcal{M}|^{a_1})|^{a_2})\ldots)|^{a_n}$.

Intuitively, the agent 'carries' the uncertainty set with him when moving forward and obtains a new uncertainty set $\mathcal{U}|^a$. Note that here we differ from [33] where the updated uncertainty set is further refined according to what the agent can observe at the new state. For conformant planning, we do not consider the observational power of the agent during the execution of a plan.

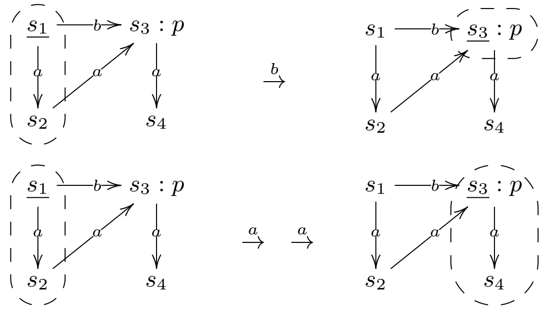
Let us call the model mentioned in the introduction $\mathcal{M}$, it is not hard to see that $\mathcal{M}|^r$ and $(\mathcal{M}|^r)|^u$ are as follows:



Thus we have:

- $\mathcal{M}, s_3 \vDash [r](\mathit{Safe} \wedge \neg K \mathit{Safe})$
- $\mathcal{M}, s_3 \vDash K[r][u](\mathit{Safe} \wedge K\mathit{Safe})$

The usual global model checking algorithm for modal logics labels the states with the subformulas that are true on the states. However, this cannot work here since the truth value of epistemic formulas on the states outside $\mathcal{U}$ is simply undefined. Moreover, the exact truth value of an epistemic formula on a state depends on 'how you get there', as the following example shows (the underlined states mark the actual states):



Let the left-hand-side model be $\mathcal{M}$ then it is clear that $\mathcal{M}|^b, s_3 \vDash Kp$ while $\mathcal{M}|^{aa}, s_3 \nvDash Kp$ thus $\mathcal{M}, s_1 \vDash \langle b\rangle Kp \wedge \langle a\rangle\langle a\rangle\neg Kp$. This shows that the truth value of an epistemic subformula w.r.t. a state in the model is somehow 'context-dependent', which requires new techniques in model checking. We will make this explicit in Section 4.3 when we discuss the model checking algorithm.

2.2 Axiomatization

Following the axioms proposed in [33], we give the following axiomatization for EAL w.r.t. our semantics:

System $\mathbb{SELA}$

Axioms		**Rules**	
TAUT	all axioms of propositional logic	MP	$\dfrac{\phi, \phi \to \psi}{\psi}$
DISTK	$K(p \to q) \to (Kp \to Kq)$	NECK	$\dfrac{\phi}{K\phi}$
DIST(a)	$[a](p \to q) \to ([a]p \to [a]q)$	NEC(a)	$\dfrac{\phi}{[a]\phi}$
T	$Kp \to p$	SUB	$\dfrac{\phi(p)}{\phi(\psi)}$
4	$Kp \to KKp$		
5	$\neg Kp \to K\neg Kp$		
PR(a)	$K[a]p \to [a]Kp$		
NM(a)	$\langle a\rangle Kp \to K[a]p$		

where a ranges over A, p, q range over P. PR($\cdot$) and NM($\cdot$) denote the axioms of *perfect recall* and *no miracles* respectively (cf. [32]).

Note that since we do not assume that the agent can observe the available actions, the axiom OBS(a) : $K\langle a\rangle\top \vee K\neg\langle a\rangle\top$ in [33] is abandoned. Due to the same reason, the axiom of no miracles is also simplified.

We show the completeness of $\mathbb{SELA}$ using a more direct proof strategy compared to the one used in [33].

THEOREM 2.1. *$\mathbb{SELA}$ is sound and strongly complete w.r.t. EAL on uncertainty maps.*

PROOF. To prove that $\mathbb{SELA}$ is sound on uncertainty maps, we need to show that all the axioms are valid and all the inference rules preserve validity. Since the uncertainty set in an UM denotes an equivalent class, axioms T, 4 and 5 are valid; due to the semantics, the validity of axioms PR($\cdot$) and NM($\cdot$) can be proved step by step; others can be proved as usual.

To prove that $\mathbb{SELA}$ is strongly complete on uncertainty maps, we only need to show that every $\mathbb{SELA}$-consistent set of formulas is satisfiable on some uncertainty map. The proof idea is that we construct an uncertainty map consisting of maximal $\mathbb{SELA}$-consistent sets (MCSs), and then with the Lindenbaum-like lemma that every $\mathbb{SELA}$-consistent set of formulas can be extended in to a MCS (we omit the proof here), we only need to prove that every formula holds on the MCS to which it belongs.

Firstly, we construct a canonical Kripke model $\mathcal{N}^c = \langle \mathcal{S}^c, \{\mathcal{R}^c_a \mid a \in \mathtt{A}\}, \mathcal{V}^c\rangle$ as follows:

- $\mathcal{S}^c$ is the set of all MCSs;
- $sR^c_a t \iff \langle a\rangle\phi \in s$ for any $\phi \in t$ (equivalently $\phi \in t$ for any $[a]\phi \in s$);
- $\mathcal{V}^c(p) = \{s \mid p \in s\}$.

Given $s \in \mathcal{S}^c$, we define $\mathcal{U}^c_s = \{u \in \mathcal{S}^c \mid K\phi \in s \text{ iff } K\phi \in u\}$, and it is obvious that $s \in \mathcal{U}^c_s$. Thus we have that for each $s \in \mathcal{S}^c$, $\mathcal{M}^c_s = \langle \mathcal{N}^c, \mathcal{U}^c_s\rangle$ is an uncertainty map, and $\mathcal{M}^c_s, s$ is a pointed uncertainty map.

Secondly, we prove the following claim.

Claim 2.1. *If $s \xrightarrow{a} t$, then we have $\mathcal{U}^c_s|^a = \mathcal{U}^c_t$.*

$\subseteq$: Assuming $v \in \mathcal{U}^c_s|^a$, we need to show $v \in \mathcal{U}^c_t$, namely we need to show that $K\phi \in v \iff K\phi \in t$. Since $v \in \mathcal{U}^c_s|^a$, we have that there is $u \in \mathcal{U}^c_s$ such that $uR^c_a v$. If $K\phi \in t$, it follows by axiom 4 that $KK\phi \in t$. Thus we have $\langle a\rangle KK\phi \in s$. By axiom $\mathtt{NM}(a)$, it follows that $K[a]K\phi \in s$. By $u \in \mathcal{U}^c_s$ and axiom $\mathtt{T}$, we have $[a]K\phi \in u$. It follows by $uR^c_a v$ that $K\phi \in v$. If $K\phi \notin t$, we have $\neg K\phi \in t$. By axiom $\mathtt{5}$, we have $K\neg K\phi \in t$. Similarly, we have $\neg K\phi \in v$. Thus we have $K\phi \notin v$.

$\supseteq$: Assuming $v \in \mathcal{U}^c_t$, we need to show $v \in \mathcal{U}^c_s|^a$, namely there is $u \in \mathcal{U}^c_s$ such that $uR^c_a v$. Let u^- be $\{K\phi \mid K\phi \in s\} \cup \{\langle a\rangle\psi \mid \psi \in v\}$. Then u^- is consistent. For suppose not, we have $\vdash K\phi_1 \wedge \dots \wedge K\phi_n \to [a]\neg\psi_1 \vee \dots \vee [a]\neg\psi_k$ for some n and k. Since $\vdash [a]\neg\psi_1 \vee \dots \vee [a]\neg\psi_k \to [a](\neg\psi_1 \vee \dots \vee \neg\psi_k)$, we have $\vdash K\phi_1 \wedge \dots \wedge K\phi_n \to [a](\neg\psi_1 \vee \dots \vee \neg\psi_k)$. By rule $\mathtt{NECK}$ and axiom $\mathtt{DISTK}$, we have $\vdash KK\phi_1 \wedge \dots \wedge KK\phi_n \to K[a](\neg\psi_1 \vee \dots \vee \neg\psi_k)$. Since $KK\phi_i \in s$ for each $1 \leq i \leq n$, we have $K[a](\neg\psi_1 \vee \dots \vee \neg\psi_k) \in s$. By axiom $\mathtt{PR}(a)$, it follows that $[a]K(\neg\psi_1 \vee \dots \vee \neg\psi_k) \in s$. It follows by $sR^c_a t$ that $K(\neg\psi_1 \vee \dots \vee \neg\psi_k) \in t$. Since $v \in \mathcal{U}^c_t$, by axiom $\mathtt{T}$, we have $\neg\psi_1 \vee \dots \vee \neg\psi_k \in v$. This is contrary with $\psi_i \in v$ for each $1 \leq i \leq k$. Thus u^- is consistent. By Lindenbaum-like Lemma, there exists a MCS u extending u^-. It follows by $u^- \subseteq u$ that $u \in \mathcal{U}^c_s$ and $uR^c_a v$. We conclude that $v \in \mathcal{U}^c_s|^a$.

Finally, we will show that $\mathcal{M}^c_s, s \vDash \phi$ iff $\phi \in s$. we prove it by induction on ϕ. Please note that the 'existence lemmas' (that $\neg[a]\phi \in s$ implies $\neg\phi \in t$ for some t such that $s \xrightarrow{a} t$ and that $\neg K\phi \in s$ implies $\neg\phi \in s'$ for some $s' \in \mathcal{U}^c_s$) also hold in the model $\mathcal{N}^c$. We only focus on the case of $[a]\phi$. With Claim 2.1, it follows that $\mathcal{M}^c_t = \mathcal{M}^c_s|^a$ if $s \xrightarrow{a} t$. Then by the induction hypothesis and the existence lemmas, it is easy to show that $\mathcal{M}^c_s, s \vDash [a]\phi$ iff $[a]\phi \in s$. $\Box$

3. AN EXTENSION OF EAL FOR CONFORMANT PLANNING

3.1 Epistemic PDL over uncertainty maps

In this section we extend the language of $\mathtt{EAL}$ with programs in propositional dynamic logic and use this extended language to express the existence of a conformant plan.

Definition 3.1 (Epistemic PDL). *The Epistemic PDL Language ($\mathtt{EPDL}$) is defined as follows:*

$$\phi ::= \top \mid p \mid \neg\phi \mid (\phi \wedge \phi) \mid [\pi]\phi \mid K\phi$$
$$\pi ::= a \mid ?\phi \mid (\pi;\pi) \mid (\pi + \pi) \mid \pi^*$$

where $p \in \mathtt{P}$, $a \in \mathtt{A}$. We use $(\!|\pi|\!)\phi$ to denote $[\pi]\phi \wedge \langle\pi\rangle\phi$, which is logically equivalent to $[\pi]\phi \wedge \langle\pi\rangle\top$. Given a finite $\mathtt{B} \subseteq \mathtt{A}$, we write $\mathtt{B}^*$ for $(\Sigma_{a\in\mathtt{B}}a)^*$, i.e., the iteration over the 'sum' of all the action symbols in $\mathtt{B}$. The size of $\mathtt{EPDL}$ formulas/programs is given by: $|[\pi]\phi| = |\pi| + |\phi|$, $|a| = 1$, $|\pi_1;\pi_2| = 1 + |\pi_1| + |\pi_2|$, $|?\phi| = |\pi^*| = 1 + |\phi|$, and $|\pi_1 + \pi_2| = 1 + |\pi_1| + |\pi_2|$.

Given any uncertainty map $\mathcal{M} = \langle \mathcal{S}, \{\mathcal{R}_a \mid a \in \mathtt{A}\}, \mathcal{V}, \mathcal{U}\rangle$, any state $s \in \mathcal{U}$, the semantics is given by a mutual induction on ϕ and π (we only show the case about $[\pi]\phi$, other cases are as in $\mathtt{EAL}$):

$$\boxed{\begin{array}{rl}
\mathcal{M}, s \vDash [\pi]\phi \Leftrightarrow & \text{for all } \mathcal{M}', s' : (\mathcal{M}, s)[\![\pi]\!](\mathcal{M}', s') \\
& \text{implies } \mathcal{M}', s' \vDash \phi \\
(\mathcal{M}, s)[\![a]\!](\mathcal{M}', s') \Leftrightarrow & \mathcal{M}' = \mathcal{M}|^a \text{ and } s \xrightarrow{a} s' \\
(\mathcal{M}, s)[\![?\psi]\!](\mathcal{M}', s') \Leftrightarrow & (\mathcal{M}', s') = (\mathcal{M}, s) \text{ and } \mathcal{M}, s \vDash \psi \\
(\mathcal{M}, s)[\![\pi_1;\pi_2]\!](\mathcal{M}', s') \Leftrightarrow & (\mathcal{M}, s)[\![\pi_1]\!] \circ [\![\pi_2]\!](\mathcal{M}', s') \\
(\mathcal{M}, s)[\![\pi_1 + \pi_2]\!](\mathcal{M}', s') \Leftrightarrow & (\mathcal{M}, s)[\![\pi_1]\!] \cup [\![\pi_2]\!](\mathcal{M}', s') \\
(\mathcal{M}, s)[\![\pi^*]\!](\mathcal{M}', s') \Leftrightarrow & (\mathcal{M}, s)[\![\pi]\!]^\star(\mathcal{M}', s')
\end{array}}$$

where $\circ, \cup, {}^\star$ at the right-hand side denote the usual composition, union and reflexive transitive closure of binary relations respectively. Clearly this semantics coincides with the semantics of $\mathtt{EAL}$ on $\mathtt{EAL}$ formulas.

Note that each program π can be viewed as a set of computation sequences, which are sequences of actions in $\mathtt{A}$ and tests with $\phi \in \mathtt{EPDL}$:

$\mathcal{L}(a) = \{a\}$
$\mathcal{L}(?\phi) = \{?\phi\}$
$\mathcal{L}(\pi;\pi') = \{\sigma\eta \mid \sigma \in \mathcal{L}(\pi) \text{ and } \eta \in \mathcal{L}(\pi')\}$
$\mathcal{L}(\pi + \pi') = \mathcal{L}(\pi) \cup \mathcal{L}(\pi')$
$\mathcal{L}(\pi^*) = \{\epsilon\} \cup \bigcup_{n>0}(\mathcal{L}(\underbrace{\pi\cdots\pi}_{n}))$ where ϵ is the empty sequence

Here are some valid formulas which are useful in our latter discussion:

$$\langle\pi;\pi'\rangle\phi \leftrightarrow \langle\pi\rangle\langle\pi'\rangle\phi$$
$$[\pi + \pi']\phi \leftrightarrow [\pi]\phi \wedge [\pi']\phi$$
$$[?\psi]\phi \leftrightarrow (\psi \to \phi)$$

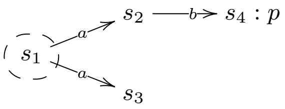
We leave the complete axiomatization of $\mathtt{EPDL}$ on uncertainty maps to future work.

3.2 Conformant planning via model checking EPDL

Definition 3.2 (Conformant planning). *Given an uncertainty map $\mathcal{M}$, a goal formula $\phi \in \mathtt{EPDL}$, and a set $\mathtt{B} \subseteq \mathtt{A}$, the conformant planning problem is to find a finite (possibly empty) sequence $\sigma = a_1a_2\cdots a_n \in \mathcal{L}(\mathtt{B}^*)$ such that for each $u \in \mathcal{U}_\mathcal{M}$ we have $\mathcal{M}, u \vDash (\!|a_1|\!)(\!|a_2|\!)\cdots(\!|a_n|\!)\phi$. The existence problem of conformant planning is to test whether such a sequence exists.*

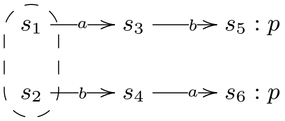
Recall that $(\!|\pi|\!)\phi$ is the shorthand of $[\pi]\phi \wedge \langle\pi\rangle\phi$. Intuitively, we want a plan which is both executable and safe w.r.t. non-deterministic actions and initial uncertainty of the agent. It is crucial to observe the difference between $(\!|a_1|\!)(\!|a_2|\!)\cdots(\!|a_n|\!)\phi$ and $(\!|a_1;a_2;\cdots;a_n|\!)\phi$ by the following example:

Example 1. *Given uncertainty map $\mathcal{M}$ depicted as follows, we have $\mathcal{M}, s_1 \vDash \llparenthesis a; b \rrparenthesis p$ but $\mathcal{M}, s_1 \nvDash \llparenthesis a \rrparenthesis \llparenthesis b \rrparenthesis p$.*



Given $\mathcal{M}$ and ϕ, to verify whether $\sigma \in \mathcal{L}(\pi)$ is a conformant plan can be formulated as the model checking problem: $\mathcal{M}, u \vDash K\llparenthesis a_1 \rrparenthesis \llparenthesis a_2 \rrparenthesis \cdots \llparenthesis a_n \rrparenthesis \phi$. On the other hand, the existence problem of a conformant plan is more complicated to formulate: it asks whether there *exists* a $\sigma \in \mathcal{L}(\mathtt{B}^*)$ such that it can be verified as a conformant plan. The simple-minded attempt would be to check whether $\mathcal{M}, u \vDash K\langle \mathtt{B}^* \rangle \phi$ holds. Despite the $\langle\cdot\rangle$-vs.-$\llparenthesis\cdot\rrparenthesis$ distinction, $K\langle \mathtt{B}^* \rangle \phi$ may hold on a model where the sequences to guarantee ϕ on different states in $\mathcal{U}_M$ are different, as the following example shows:

Example 2. *Given uncertainty map $\mathcal{M}$ depicted as follows, let the goal formula be p and $\mathtt{B} = \{a, b\}$. We have $\mathcal{M}, s_1 \vDash K\langle \mathtt{B}^* \rangle p$, but there is no solution to this conformant planning problem.*



The right formula to check for the existence of a conformant plan w.r.t. $\mathtt{B} \subseteq \mathtt{A}$ and $\phi \in \mathtt{EPDL}$ is:

$$\theta_{\mathtt{B},\phi} = \langle (\Sigma_{a \in \mathtt{B}} (?K\langle a \rangle \top; a))^* \rangle K\phi.$$

For example, if $\mathtt{B} = \{a_1, a_2\}$ then $\theta_{\mathtt{B},\phi} = \langle ((?K\langle a_1 \rangle \top; a_1) + (?K\langle a_2 \rangle \top; a_2))^* \rangle K\phi$. Intuitively, the confrmant plan consists of actions that are always executable given the uncertainty of the agent (guaranteed by the guard $K\langle a \rangle \top$). In the end the plan should also make sure that ϕ must hold given the uncertainty of the agent (guaranteed by $K\phi$). In the following, we will prove that this formula is indeed correct.

First, we observe that the rule of substitution of equivalents is valid ($\phi(\psi/\chi)$ is obtained by replacing any occurrence of χ by ψ, similar for $\llbracket \pi(\psi/\chi) \rrbracket$):

Proposition 3.1. *If $\vDash \psi \leftrightarrow \chi$, then:*

(1) $\vDash \phi \leftrightarrow \phi(\psi/\chi)$;

(2) $\llbracket \pi \rrbracket = \llbracket \pi(\psi/\chi) \rrbracket$.

Proposition 3.2. *$\vDash K\llparenthesis a \rrparenthesis \phi \leftrightarrow \langle ?K\langle a \rangle \top; a \rangle K\phi$*

Proof. Since $\vDash K\llparenthesis a \rrparenthesis \phi \leftrightarrow (K[a]\phi \wedge K\langle a \rangle \phi)$ and $\vDash (K\langle a \rangle \top \wedge \langle a \rangle K\phi) \leftrightarrow \langle ?K\langle a \rangle \top; a \rangle K\phi$, we only need to show that $\vDash (K[a]\phi \wedge K\langle a \rangle \phi) \leftrightarrow (K\langle a \rangle \top \wedge \langle a \rangle K\phi)$.

Left to right:
(L1) $\vDash K[a]\phi \rightarrow [a]K\phi$, by validity of Axiom $\mathtt{PR}(a)$
(L2) $\vDash K\langle a \rangle \phi \rightarrow \langle a \rangle \top \wedge K\langle a \rangle \top$, by semantics
(L3) $\vDash \langle a \rangle \top \wedge [a]K\phi \rightarrow \langle a \rangle K\phi$, by semantics
(L4) $\vDash K[a]\phi \wedge K\langle a \rangle \phi \rightarrow K\langle a \rangle \top \wedge \langle a \rangle K\phi$, by (L1)-(L3)
Right to left:
(R1) $\vDash \langle a \rangle K\phi \rightarrow K[a]\phi$, by validity of Axiom $\mathtt{NM}(a)$
(R2) $\vDash K[a]\phi \wedge K\langle a \rangle \top \rightarrow K\langle a \rangle \phi$, by semantics
(R3) $\vDash K\langle a \rangle \top \wedge \langle a \rangle K\phi \rightarrow K[a]\phi \wedge K\langle a \rangle \phi$, by R(1)-R(2) □

Lemma 3.1. *For any $a_1 a_2 \cdots a_n \in \mathcal{L}(\mathtt{A}^*)$:*

$$\vDash K\llparenthesis a_1 \rrparenthesis \llparenthesis a_2 \rrparenthesis \cdots \llparenthesis a_n \rrparenthesis \phi \leftrightarrow \langle ?K\langle a_1 \rangle \top; a_1; \ldots; ?K\langle a_n \rangle \top; a_n \rangle K\phi$$

Proof. It is trivial when $n = 0$ (i.e., the sequence is ϵ), since the claim then boils down to $K\phi \leftrightarrow K\phi$. We prove the non-trivial cases by induction on $n \geq 1$. When $n = 1$, it follows from Proposition 3.2. Now, as the induction hypothesis, we assume that:

$$\vDash K\llparenthesis a_1 \rrparenthesis \llparenthesis a_2 \rrparenthesis \cdots \llparenthesis a_k \rrparenthesis \phi \leftrightarrow \langle ?K\langle a_1 \rangle \top; a_1; \ldots; ?K\langle a_k \rangle \top; a_k \rangle K\phi.$$

We need to show:

$$\begin{aligned}&\vDash K\llparenthesis a_1 \rrparenthesis \llparenthesis a_2 \rrparenthesis \cdots \llparenthesis a_{k+1} \rrparenthesis \phi \leftrightarrow \\ &\quad \langle ?K\langle a_1 \rangle \top; a_1; \ldots; ?K\langle a_{k+1} \rangle \top; a_{k+1} \rangle K\phi.\end{aligned}$$

By IH,

$$\begin{aligned}&\vDash K\llparenthesis a_1 \rrparenthesis \llparenthesis a_2 \rrparenthesis \cdots \llparenthesis a_{k+1} \rrparenthesis \phi \leftrightarrow \\ &\quad \langle ?K\langle a_1 \rangle \top; a_1; \ldots; ?K\langle a_k \rangle \top; a_k \rangle K\llparenthesis a_{k+1} \rrparenthesis \phi. \qquad (1)\end{aligned}$$

Due to Propositions 3.1 and 3.2, we have

$$\begin{aligned}&\vDash \langle ?K\langle a_1 \rangle \top; a_1; \ldots; ?K\langle a_k \rangle \top; a_k \rangle K\llparenthesis a_{k+1} \rrparenthesis \phi \leftrightarrow \\ &\langle ?K\langle a_1 \rangle \top; a_1; \ldots; ?K\langle a_n \rangle \top; a_k \rangle \langle ?K\langle a_{k+1} \rangle \top; a_{k+1} \rangle K\phi. \; (2)\end{aligned}$$

The conclusion is immediate by combining (1) and (2). □

The following theorem follows from the above lemma.

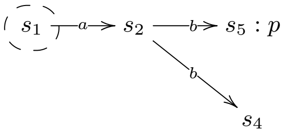
Theorem 3.1. *Given a pointed uncertainty map $\mathcal{M}, s$, an $\mathtt{EPDL}$ formula ϕ and a set $\mathtt{B} \subseteq \mathtt{A}$, the following two are equivalent:*

(1) There is a $\sigma = a_1 \ldots a_n \in \mathcal{L}(\mathtt{B}^*)$ such that $\mathcal{M}, s \vDash K\llparenthesis a_1 \rrparenthesis \llparenthesis a_2 \rrparenthesis \cdots \llparenthesis a_n \rrparenthesis \phi$;

(2) $\mathcal{M}, s \vDash \langle (\Sigma_{a \in \mathtt{B}} (?K\langle a \rangle \top; a))^* \rangle K\phi$.

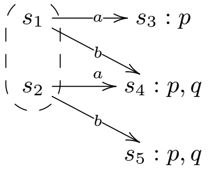
We would like to emphasise that the K operator right before ϕ in the definition of $\theta_{\mathtt{B},\phi}$ cannot be omitted, as demonstrated by the following example:

Example 3. *Given uncertainty map $\mathcal{M}$ depicted as follows, let the goal formula be p. As we can see, there is no solution to this conformant planning problem. Indeed $\mathcal{M}, s_1 \nvDash \langle (\Sigma_{a \in \mathtt{B}} (?K\langle a \rangle \top; a))^* \rangle Kp$ with $\mathtt{B} = \{a, b\}$, but we could have $\mathcal{M}, s_1 \vDash \langle (\Sigma_{a \in \mathtt{B}} (?K\langle a \rangle \top; a))^* \rangle p$.*



We close this section with an example about planning with both positive and negative epistemic goals (the agent should know something, but not too much).

Example 4. *Given uncertainty map $\mathcal{M}$ depicted as follows, let the goal be Kp then both a and b are conformant plans. If the goal is $Kp \wedge \neg Kq$, only a is a good plan.*



4. MODEL CHECKING EPDL: COMPLEXITY AND ALGORITHMS

In this section, we first focus on the model checking problem of the following star-free fragment of `EPDL` (call it $\texttt{EPDL}^-$):

$$\phi ::= \top \mid p \mid \neg\phi \mid (\phi \wedge \phi) \mid [\pi]\phi \mid K\phi$$
$$\pi ::= a \mid ?\phi \mid (\pi;\pi) \mid (\pi + \pi)$$

We will show that model checking $\texttt{EPDL}^-$ is PSPACE-complete. In particular, the upper bound is shown by making use of an alternative context-dependent semantics. Then we give an EXPTIME algorithm for the model checking problem of the full `EPDL` inspired by another alternative semantics based on 2-dimensional models. Finally we give a PSPACE algorithm for the conformant planning problem in `EPDL`. Note that throughout this section, we focus on uncertainty maps with finitely many states and assume $\mathcal{R}_a = \emptyset$ for *co-finitely* many $a \in \texttt{A}$.

4.1 Complexity of model checking EPDL⁻

4.1.1 Lower Bound

To show the PSPACE lower bound, we provide a polynomial reduction of QBF (*quantified Boolean formula*) truth testing to the model checking problem of $\texttt{EPDL}^-$. Note that to determine whether a given QBF (even in prenex normal form based on a conjunctive normal form) is true or not is known to be PSPACE-complete [29]. Our method is inspired by [27] which discusses the complexity of model checking temporal logics with past operators. Surprisingly, we can use the uncertainty sets to encode the 'past' and use the dual of the knowledge operator to 'go back' to the past. This intuitive idea will become more clear in the proof.

QBF formulas are $Q_1x_1Q_2x_2\dots Q_nx_n\phi(x_1,\dots,x_n)$ where:

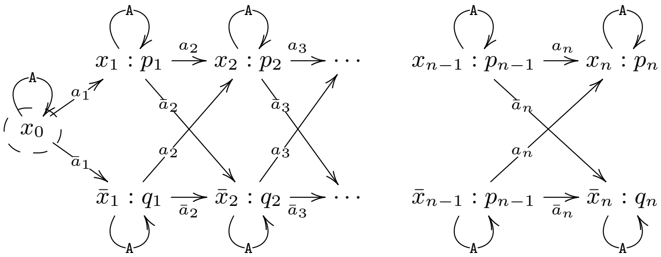
- For $1 \le n \le n, Q_i$ is $\exists$ if i is odd, and Q_i is $\forall$ if i is even.
- ϕ is a propositional formula in CNF based on variables $x_1,\dots,x_n$,

For each such QBF α with n variables, we need to find a pointed model $\mathcal{M}_n, x_0$ and a formula θ_α such that α is true iff $\mathcal{M}_n, x_0 \vDash \theta_\alpha$. The model $\mathcal{M}_n$ is defined below.

DEFINITION 4.1. *Let* $\texttt{A} = \{a_i, \bar{a}_i \mid i \ge 1\}$ *and* $\texttt{P} = \{p_k, q_k \mid k \ge 1\}$*, the uncertainty map* $\mathcal{M}_n = \langle \mathcal{S}, \{\mathcal{R}_a \mid a \in \texttt{A}\}, \mathcal{V}, \mathcal{U}\rangle$ *is defined as:*

- $\mathcal{S} = \{x_0\} \cup \{x_i \mid 1 \le i \le n\} \cup \{\bar{x}_i \mid 1 \le i \le n\}$
- $\mathcal{V}(x_0) = \emptyset$*, and* $\mathcal{V}(x_i) = \{p_i\}, \mathcal{V}(\bar{x}_i) = \{q_i\}$ *for* $1 \le i \le n$.
- $\xrightarrow{a_i} = \{(s,s) \mid s \in \mathcal{S}\} \cup \{(x_{i-1}, x_i), (\bar{x}_{i-1}, x_i)\}$
- $\xrightarrow{\bar{a}_i} = \{(s,s) \mid s \in \mathcal{S}\} \cup \{(x_{i-1}, \bar{x}_i), (\bar{x}_{i-1}, \bar{x}_i)\}$
- $\mathcal{U} = \{x_0\}$

$|\mathcal{M}_n|$ is linear in n and can be depicted as the following:



Given $\alpha = Q_1x_1Q_2x_2\dots Q_nx_n\phi(x_1,\dots,x_n)$, the formula θ_α is defined as

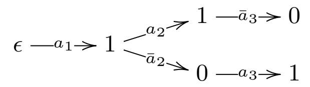
$$QT_1 \cdots QT_n\psi(\hat{K}p_1, \cdots, \hat{K}p_n, \hat{K}q_1, \cdots, \hat{K}q_n)$$

where QT_i is $\langle(a_i + \bar{a}_i); ?(p_i \vee q_i)\rangle$ if i is odd and QT_i is $[(a_i + \bar{a}_i); ?(p_i \vee q_i)]$ if i is even, and ψ is obtained from $\phi(x_1,\dots,x_n)$ by replacing each x_i with $\hat{K}p_i$ and $\neg x_i$ with $\hat{K}q_i$.

To ease the latter proof, we first define the valuation tree below.

DEFINITION 4.2 (V-TREE). *A V-tree τ is a rooted tree such that 1) each node is 0 or 1 (except the root ϵ); 2) each internal node in an even level has only one successor; 3) each internal node in an odd level has two successors: one is 0 and the other one is 1; 4) each edge to node 0 of level i is labelled $\bar{a}_i$; 5) each edge to node 1 of level i is labelled a_i. Given a V-tree with depth n, a path σ is a sequence of $A_1 \dots A_n$ where $A_i = a_i$ or $A_i = \bar{a}_i$. A path σ can also be seen as a valuation assignment for $x_1,\dots,x_n$ with the convention that $\sigma(x_i) = 1$ if a_i occurs in σ and $\sigma(x_i) = 0$ if $\bar{a}_i$ occurs in σ. Let path(τ) be the set of all paths of τ.*

As an example, a V-tree τ can be depicted as below:



It is not hard to see the following:

PROPOSITION 4.1. *For each $1 \le i \le n$, we have $\alpha = Q_1x_1\dots Q_ix_iQ_{i+1}x_{i+1}\dots Q_nx_n\phi$ is true iff there exists a V-tree τ with depth i such that $\sigma(Q_{i+1}x_{i+1}\dots Q_nx_n\phi) = 1$ for each $\sigma \in path(\tau)$ (σ is viewed as a valuation).*

Now let us see the update result of running a path $\sigma \in path(\tau)$ on $\mathcal{M}_n$. Due to the lack of space, we omit the proofs of the following two propositions.

PROPOSITION 4.2. *Given $\mathcal{M}_n$, let $\sigma = A_1 \dots A_i$ ($1 \le i \le n$) be a sequence of actions such that $A_k = a_k$ or $A_k = \bar{a}_k$ for each $1 \le k \le i$, then we have $\mathcal{U}|^\sigma = \{x_0, X_1, \dots, X_i\}$ where $X_k = x_k$ if $A_k = a_k$ else $X_k = \bar{x}_k$ for each $1 \le k \le i$.*

Given $\sigma = A_1 \dots A_n$ where A_i is a_i or $\bar{a}_i$ for each $1 \le i \le n$, let $g(\sigma) = x_n$ if $A_n = a_n$ and $g(\sigma) = \bar{x}_n$ if $A_n = \bar{a}_n$. By Proposition 4.2, we always have $g(\sigma) \in \mathcal{U}_{\mathcal{M}_k}|^\sigma$ with $k > n$. Thus given $\mathcal{M}_k$ and $\sigma = A_1 \dots A_n$ and $k > n$, $\mathcal{M}_k|^\sigma, g(\sigma)$ is a pointed uncertainty map.

PROPOSITION 4.3. *For each $1 \le i \le n$, we have $\mathcal{M}_k, x_0 \vDash QT_1 \dots QT_iQT_{i+1} \dots QT_n\psi$ iff there exists a V-tree τ with depth i such that $\mathcal{M}_k|^\sigma, g(\sigma) \vDash QT_{i+1} \dots QT_n\psi$ for each $\sigma \in path(\tau)$, where $k > n$ and $g(\sigma)$ is the state corresponds to the last edge of σ, e.g., $g(a_1\bar{a}_2) = \bar{x}_2$.*

Theorem 4.1. *The following two are equivalent:*

- $\alpha = Q_1x_1Q_2x_2\dots Q_nx_n\phi(x_1,\dots,x_n)$ *is true*
- $\mathcal{M}_n, x_0 \vDash QT_1\cdots QT_n\psi(\hat{K}p_1\cdots\hat{K}p_n, \hat{K}q_1\cdots\hat{K}q_n)$ *in which* ψ *is obtained from* ϕ *by replacing each* x_i *with* $\hat{K}p_i$ *and* $\neg x_i$ *with* $\hat{K}q_i$.

Proof. By Propositions 4.1 and 4.3, we only need to show that given V-tree τ with depth n, $\sigma(\phi) = 1$ if and only if $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \psi$ for each $\sigma \in path(\tau)$. Since ϕ is in CNF, ψ is also in CNF-like form obtained by replacing each x_i with $\hat{K}p_i$ and each $\neg x_i$ with $\hat{K}q_i$ for $1 \leq i \leq n$. Thus we only need to show that $\sigma(x_i) = 1$ iff $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \hat{K}p_i$ and $\sigma(\neg x_i) = 1$ iff $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \hat{K}q_i$. Since $\sigma(x_i) = 1$ iff $\sigma(\neg x_i) = 0$, we only need to show that $\sigma(x_i) = 1$ iff $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \hat{K}p_i$ and $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \hat{K}p_i$ iff $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \neg\hat{K}q_i$. By the definition of τ, we know that $\sigma = A_1\dots A_n$ where A_i is a_i or $\bar{a}_i$ for each $1 \leq i \leq n$.

Firstly, we will show that $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \hat{K}p_i$ if and only if $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \neg\hat{K}q_i$. To verify the right-to-left direction, if $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \hat{K}p_i$, it follows by the definition of $\mathcal{M}_n$ that $x_i \in \mathcal{U}|^{\sigma}$. Then it must be the case that a_i occurs in σ. Suppose not, $\bar{a}_i$ occurs in σ. It follows by Proposition 4.2, $\mathcal{U}|^{\sigma} = \{x_0, X_1, \dots, X_{i-1}, \bar{x}_i, X_{i+1}, \dots, X_n\}$. This is contrary with $x_i \in \mathcal{U}|^{\sigma}$. Thus it must be that a_i occurs in σ. It follows by Proposition 4.2 that $\mathcal{U}|^{\sigma} = \{x_0, X_1, \dots, X_{i-1}, x_i, X_{i+1}, \dots, X_n\}$. Thus $\bar{x}_i \notin \mathcal{U}|^{\sigma}$. By the definition of $\mathcal{M}_n$ and the semantics, we have $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \neg\hat{K}q_i$. To verify the left-to-right direction, $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \neg\hat{K}q_i$ implies that $\bar{x}_i \notin \mathcal{U}|^{\sigma}$. For the similar reason as above, it must be the case that $\bar{a}_i$ does not occur in σ. Thus we have that a_i occurs in σ. It follows by Proposition 4.2 that $x_i \in \mathcal{U}|^{\sigma}$. Thus we have $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \hat{K}p_i$.

Next we will show that $\sigma(x_i) = 1$ iff $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \hat{K}p_i$. To verify the right-to-left direction, $\sigma(x_i) = 1$ implies that $A_i = a_i$. It follows by Proposition 4.2 that $x_i \in \mathcal{U}|^{\sigma}$. Thus we have $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \hat{K}p_i$. To verify the left-to-right direction, we will show that $\sigma(x_i) = 0$ implies $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \hat{K}q_i$. It follows by the definition of $\sigma(x_i) = 0$ that $A_i = \bar{a}_i$. It follows by Proposition 4.2 that $\bar{x}_i \in \mathcal{U}|^{\sigma}$. Thus we have $\mathcal{M}_n|^{\sigma}, g(\sigma) \vDash \hat{K}q_i$. □

This gives us the desired lower bound:

Theorem 4.2. *The model checking problem for* EPDL$^-$ *is* Pspace*-hard.*

4.1.2 Upper Bound

In this section we give a non-trivial model checking algorithm for EPDL$^-$ inspired by an equivalent semantics.

As we mentioned earlier, the semantics of EPDL is 'context-dependent': reaching the same state through different paths may affect the truth value of an epistemic subformula. This means that the usual global model checking algorithm for modal logics may not work here. In order to establish the upper bound, we first give the following equivalent semantics to EPDL$^-$ which makes the context dependency explicit in order to facilitate a local model checking algorithm. The idea is to keep the model intact but record the scope of action modalities in order to compute the right uncertainty set for epistemic subformulas. Similar idea appeared in [32] to give an alternative semantics of public announcement logic.

Definition 4.3. *Given an uncertainty map* $\mathcal{M} = \langle \mathcal{S}, \{\mathcal{R}_a \mid a \in \mathtt{A}\}, \mathcal{V}, \mathcal{U}\rangle$ *and any state* $s \in \mathcal{S}$, *the satisfaction relation* $\Vdash$ *is defined using the auxiliary satisfaction relation* $\Vdash_\sigma$ *and auxiliary relation* $\overset{\omega}{\to}_{\sigma}$, *where* σ *is a finite (possibly empty) sequence of actions in* A*:*

$\mathcal{M}, s \Vdash \phi$	$\Leftrightarrow \mathcal{M}, s \Vdash_\epsilon \phi$
$\mathcal{M}, s \Vdash_\sigma \top$	$\Leftrightarrow$ *always*
$\mathcal{M}, s \Vdash_\sigma p$	$\Leftrightarrow p \in \mathcal{V}(s)$
$\mathcal{M}, s \Vdash_\sigma \neg\phi$	$\Leftrightarrow \mathcal{M}, s \nVdash_\sigma \phi$
$\mathcal{M}, s \Vdash_\sigma \phi \wedge \psi$	$\Leftrightarrow \mathcal{M}, s \Vdash_\sigma \phi$ *and* $\mathcal{M}, s \Vdash_\sigma \psi$
$\mathcal{M}, s \Vdash_\sigma K\phi$	$\Leftrightarrow$ *for all* $v \in \mathcal{U}|^{\sigma} : \mathcal{M}, v \Vdash_\sigma \phi$
$\mathcal{M}, s \Vdash_\sigma \langle\pi\rangle\phi$	$\Leftrightarrow$ *there exists* $\omega \in \mathcal{L}(\pi)$ *and* $t \in \mathcal{S}$ *such that* $s \overset{\omega}{\to}_{\sigma} t$ *and* $\mathcal{M}, t \Vdash_{\sigma r(\omega)} \phi$
$s \overset{\epsilon}{\to}_{\sigma} t$	$\Leftrightarrow s = t$
$s \overset{(a\omega')}{\to}_{\sigma} t$	$\Leftrightarrow$ *there exists* s' *such that* $s \overset{a}{\to} s'$ *and* $s' \overset{\omega'}{\to}_{(\sigma a)} t$
$s \overset{(?\phi\omega')}{\to}_{\sigma} t$	$\Leftrightarrow \mathcal{M}, s \Vdash_\sigma \phi$ *and* $s \overset{\omega'}{\to}_{\sigma} t$

where $r(\omega)$ *is the sequence of actions obtained by eliminating all the tests in* ω.

Note that ω in the above definition is a computation sequence, i.e., a finite sequence of actions and EPDL$^-$-tests, while σ is a test-free sequence of actions.

The following can be proved by induction on η:

Proposition 4.4. *Given an uncertainty map* $\mathcal{M}$ *and sequences of actions and tests* η, ω, ω' *such that* $\eta = \omega\omega'$, *we have* $(s,t) \in \overset{\eta}{\to}_{\sigma}$ *iff* $(s,t) \in \overset{\omega}{\to}_{\sigma} \circ \overset{\omega'}{\to}_{\sigma r(\omega)}$ *for any sequence of actions* σ.

Proof. We prove it by induction on $|\eta|$. If $|\eta| \leq 2$, it is obvious by the definition. If $|\eta| > 2$, there are two cases, that is, $\eta = a\eta'$ or $\eta =?\phi\eta'$.

Case $\eta = a\eta'$: We have $\omega = a\omega''$ for some initial segment ω'' of η', and $(s,t) \in \overset{(a\eta')}{\to}_{\sigma}$ iff there exists s' such that $s \overset{a}{\to} s'$ and $(s',t) \in \overset{\eta'}{\to}_{\sigma a}$. By IH, we have $\overset{\eta'}{\to}_{\sigma a} = \overset{\omega''}{\to}_{\sigma a} \circ \overset{\omega'}{\to}_{\sigma a r(\omega'')}$. Thus we have $(s',t) \in \overset{\eta'}{\to}_{\sigma a}$ iff there exists t' such that $(s',t') \in \overset{\omega''}{\to}_{\sigma a}$ and $(t',t) \in \overset{\omega'}{\to}_{\sigma a r(\omega'')}$. By definition, we have that $s \overset{a}{\to} s'$ and $(s',t') \in \overset{\omega''}{\to}_{\sigma a}$ iff $(s,t') \in \overset{a\omega''}{\to}_{\sigma}$. Thus we have $(s,t) \in \overset{a\omega''}{\to}_{\sigma} \circ \overset{\omega'}{\to}_{\sigma a r(\omega'')}$, namely $(s,t) \in \overset{\omega}{\to}_{\sigma} \circ \overset{\omega'}{\to}_{\sigma r(\omega)}$.

Case $\eta =?\phi\eta'$: We have $\omega =?\phi\omega''$ for some initial segment ω'' of η', and $(s,t) \in \overset{(?\phi\eta')}{\to}_{\sigma}$ iff $\mathcal{M}, s \Vdash_\sigma \phi$ and $s \overset{\eta'}{\to}_{\sigma} t$. By IH, we have $s \overset{\eta'}{\to}_{\sigma} t$ iff $(s,t) \in \overset{\omega''}{\to}_{\sigma} \circ \overset{\omega'}{\to}_{\sigma r(\omega'')}$. Thus we have there exists s' such that $(s,s') \in \overset{\omega''}{\to}_{\sigma}$ and $(s',t) \in \overset{\omega'}{\to}_{\sigma r(\omega'')}$. This follows that $(s,s') \in \overset{(?\phi\omega'')}{\to}_{\sigma}$, and $(s,t) \in \overset{(?\phi\omega'')}{\to}_{\sigma} \circ \overset{\omega'}{\to}_{\sigma r(?\phi\omega'')}$, namely $(s,t) \in \overset{\omega}{\to}_{\sigma} \circ \overset{\omega'}{\to}_{\sigma r(\omega)}$. □

In the following we show that $\Vdash$ coincides with $\vDash$.

Theorem 4.3. *Given an uncertainty map* $\mathcal{M}$ *and an action sequence* σ, *if* $\mathcal{U}|^{\sigma} \neq \emptyset$, *we have that for each* $s \in \mathcal{U}|^{\sigma}$,

(i) $\mathcal{M}|^{\sigma}, s[\![\pi]\!]\mathcal{M}', s'$ *iff there exists* $\omega \in \mathcal{L}(\pi)$ *such that* $\mathcal{M}' = \mathcal{M}|^{\sigma r(\omega)}$ *and* $s \overset{\omega}{\to}_{\sigma} s'$,

(ii) $\mathcal{M}|^{\sigma}, s \vDash \phi$ *iff* $\mathcal{M}, s \Vdash_{\sigma} \phi$.

Proof. The proof is by simultaneous induction on π and ϕ (due to the test actions). For (i), we will only focus on the case of $\pi_1;\pi_2$; the other cases are straightforward.

Case $\pi_1;\pi_2$: We only show the direction from left to right; the other direction is similar. It follows by assumption that there is pointed uncertainty map $\mathcal{N}, t$ such that $\mathcal{M}|^{\sigma}, s[\![\pi_1]\!]\mathcal{N}, t$ and $\mathcal{N}, t[\![\pi_2]\!]\mathcal{M}', s'$. By IH, we have that there exists $\omega \in \mathcal{L}(\pi_1)$ such that $\mathcal{N} = \mathcal{M}|^{\sigma r(\omega)}$ and $s \xrightarrow{\omega_\sigma} t$. Since $\mathcal{N}, t$ is a pointed uncertainty map and $\mathcal{N} = \mathcal{M}|^{\sigma r(\omega)}$, we have $t \in \mathcal{U}|^{\sigma r(\omega)}$. By IH and $\mathcal{M}|^{\sigma r(\omega)}, t[\![\pi_2]\!]\mathcal{M}', s'$, we have that there exists $\omega' \in \mathcal{L}(\pi_2)$ such that $\mathcal{M}|^{\sigma r(\omega) r(\omega')} = \mathcal{M}|^{\sigma r(\omega\omega')} = \mathcal{M}'$ and $t \xrightarrow{\omega'_{\sigma r(\omega)}} s'$. By Proposition 4.4, it follows that $\omega\omega' \in \mathcal{L}(\pi_1;\pi_2)$ and $s \xrightarrow{(\omega\omega')_\sigma} s'$.

For (ii), we will focus on the case of $\langle\pi\rangle\phi$; the other cases are straightforward.

Case $\langle\pi\rangle\phi$: We have $\mathcal{M}|^{\sigma}, s \vDash \langle\pi\rangle\phi$ if and only if there is pointed uncertainty map $\mathcal{M}', s'$ such that $\mathcal{M}|^{\sigma}, s[\![\pi]\!]\mathcal{M}', s'$ and $\mathcal{M}', s' \vDash \phi$. By (i), it follows that $\mathcal{M}|^{\sigma}, s[\![\pi]\!]\mathcal{M}', s'$ iff there exists $\omega \in \mathcal{L}(\pi)$ such that $\mathcal{M}' = \mathcal{M}|^{\sigma r(\omega)}$ and $s \xrightarrow{\omega_\sigma} s'$. By IH, it follows that $\mathcal{M}|^{\sigma r(\omega)}, s' \vDash \phi$ iff $\mathcal{M}, s' \Vdash_{\sigma r(\omega)} \phi$. Thus we have $\mathcal{M}, s \Vdash \langle\pi\rangle\phi$. □

Let σ be ϵ, we have the equivalence of $\Vdash$ and $\vDash$.

Corollary 4.1. *Given pointed uncertainty map* $\mathcal{M}, s$, *we have* $\mathcal{M}, s \vDash \phi$ *iff* $\mathcal{M}, s \Vdash \phi$ *for each* $\phi \in$ *EPDL*$^{-}$.

This alternative semantics induces a natural algorithm to compute the truth value of an EPDL^{-} formula w.r.t. to a pointed uncertainty map. The idea is to recursively call a function $MC(\mathcal{M}, s, \sigma, \phi)$ which returns the truth value of a subformula ϕ on state s given the context of σ while keeping $\mathcal{M}$ intact. Note that, we do not need to compute all the $MC(\mathcal{M}, s, \sigma, \phi)$ for each σ and each subformula ϕ. The only tricky part comes when evaluating $\langle\pi\rangle\phi$ formulas since it is too space consuming to compute the whole set of $\mathcal{L}(\pi)$ in the search of the right ω. Instead, we can generate one by one in some lexicographical order all the possible sequences up to a bound based on the atomic actions and tests occurring in the formula, and then test whether it belongs to the program π. Note that in this way, we can use the space repeatedly, and the membership testing of $\mathcal{L}(\pi)$ is not expensive (NLOGspace-complete according to [19]).

In the appendix we present three algorithms based on matrix representation of the model: Algorithm 1 computes the uncertainty set $\mathcal{U}|^{\sigma}$; Algorithm 2 computes $\xrightarrow{\omega_\sigma}$ and Algorithm 3 is the main model checking algorithm. Note that Algorithms 2 and 3 involve mutual recursion of each other due to the tests in programs. However, the depth of the recursion is bounded by the length of the formula, and for each call polynomial space suffices. The detailed algorithms and complexity analysis can be found in the appendix. It is not hard to show the following (based on Theorem 4.2)

Theorem 4.4 (Upper bound). *The model checking problem of EPDL*$^{-}$ *is in* Pspace. *Thus it is* Pspace*-complete.*

4.2 Upper Bounds for model checking EPDL

In this section, we give an EXPtime model checking method for the full EPDL via model checking EPDL over two-dimensional models with both epistemic and action relations. Let us first define such models.

Definition 4.4 (Epistemic Temporal Structure). *An Epistemic Temporal Structure (ETS) is a Kripke model with both epistemic and action relations. Formally, an ETS model* $\mathfrak{M}$ *is a tuple* $\langle \mathcal{S}, \{\mathcal{R}_a \mid a \in \mathtt{A}\}, \sim, \mathcal{V}\rangle$, *where* $\mathcal{R}_a$ *is a binary relation on* $\mathcal{S}$, $\sim$ *is an equivalence relation on* $\mathcal{S}$ *and* $\mathcal{V}: \mathcal{S} \to 2^{\mathtt{P}}$ *is a valuation function.*

Now we define an alternative semantics of EPDL over ETSs.[6]

Definition 4.5 (ETS Semantics). *Given any ETS model* $\mathfrak{M} = \langle \mathcal{S}, \{\mathcal{R}_a \mid a \in \mathtt{A}\}, \sim, \mathcal{V}\rangle$ *and any state* $s \in \mathcal{S}$, *the satisfaction relation for EPDL formulas is defined as follows (the Boolean cases are as in the standard modal logic):*

$$\boxed{\begin{array}{ll}
\mathfrak{M}, s \Vdash K\phi & \Leftrightarrow \forall u \in \mathcal{S} : s \sim u \text{ implies } \mathfrak{M}, u \Vdash \phi \\
\mathfrak{M}, s \Vdash [\pi]\phi & \Leftrightarrow \forall t \in S : s \xrightarrow{\pi} t \text{ implies } \mathfrak{M}, t \Vdash \phi \\
\xrightarrow{a} & = \mathcal{R}_a \\
\xrightarrow{?\phi} & = \{(s,s) \mid \mathfrak{M}, s \Vdash \phi\} \\
\xrightarrow{\pi_1;\pi_2} & = \xrightarrow{\pi_1} \circ \xrightarrow{\pi_2} \\
\xrightarrow{\pi_1+\pi_2} & = \xrightarrow{\pi_1} \cup \xrightarrow{\pi_2} \\
\xrightarrow{\pi^*} & = (\xrightarrow{\pi})^{\star}
\end{array}}$$

where $\circ, \cup, {}^{\star}$ *at right-hand side denote the usual composition, union and reflexive transitive closure of binary relations respectively.*

We can turn a Kripke model without the epistemic relation into an ETS model by essentially considering all the possible uncertainty sets.

Definition 4.6. *Given any Kripke model* $\mathcal{M} = \langle \mathcal{S}, \{\mathcal{R}_a \mid a \in \mathtt{A}\}, \mathcal{V}\rangle$, *we define the ETS model* $\mathcal{M}^{\bullet}$ *as follows:*

$$\begin{array}{ll}
\mathcal{S}^{\bullet} & = \{s_{\Gamma} \mid s \in \mathcal{S}, \Gamma \in 2^{\mathcal{S}}, s \in \Gamma\} \\
\mathcal{R}^{\bullet}_a & = \{(s_{\Gamma}, t_{\Delta}) \mid s \xrightarrow{a} t, \Delta = \Gamma|^{a}\} \\
\sim^{\bullet} & = \{(s_{\Gamma}, t_{\Delta}) \mid \Gamma = \Delta\} \\
\mathcal{V}^{\bullet}(s_{\Gamma}) & = \mathcal{V}(s)
\end{array}$$

where $\Gamma|^{a} = \{t \in \mathcal{S} \mid \exists s \in \Gamma$ *such that* $s \xrightarrow{a} t\}$. *For any Kripke model* $\mathcal{M}$ *and any* $\Gamma \in 2^{\mathcal{S}} \backslash \{\emptyset\}$, *let* $\mathcal{M}^{\Gamma}$ *be the uncertainty map* $\langle \mathcal{M}, \Gamma\rangle$.

Note that each s_{Γ} can be viewed as an uncertainty set (Γ) with a designated state (s), and the definition of $\mathcal{R}_a$ captures the update in the $\vDash$ semantics of EPDL, and $\mathcal{M}^{\bullet}$ unravels all the updates in a whole picture. Note that the size of $\mathcal{M}^{\bullet}$ is $|\mathcal{S}| \cdot 2^{|\mathcal{S}|-1}$ where $\mathcal{S}$ is the set of states of $\mathcal{M}$.

Now we can show that $\vDash$ and $\Vdash$ coincide w.r.t. uncertainty map $\mathcal{M}^{\Gamma}$ and ETS model $\mathcal{M}^{\bullet}$ (the proofs are omitted due to the lack of space).

Proposition 4.5. *Given any map* $\mathcal{M}$, *we have*

(i) $\mathcal{M}^{\Gamma}, s[\![\pi]\!]\mathcal{M}^{\Delta}, t$ *iff* $s_{\Gamma} \xrightarrow{\pi} t_{\Delta}$ *in* $\mathcal{M}^{\bullet}$;[7]

(ii) $\mathcal{M}^{\Gamma}, s \vDash \phi$ *iff* $\mathcal{M}^{\bullet}, s_{\Gamma} \Vdash \phi$.

Corollary 4.2. *Given an uncertainty map* $\mathcal{M} = \langle \mathcal{N}, \mathcal{U}\rangle$ *and* $s \in \mathcal{U}$, *we have* $\mathcal{M}, s \vDash \phi$ *iff* $\mathcal{N}^{\bullet}, s_{\mathcal{U}} \Vdash \phi$.

Based on the above corollary we can have a model checking method via model checking EPDL over ETS models.

[6] Here we abuse the notation $\Vdash$ to denote the new semantics. Note that it is different from the alternative semantics in the previous section.

[7] Cf. the definition of $\xrightarrow{\pi}$ in Def. 4.5.

PROPOSITION 4.6. *The model checking problem of* EPDL *on uncertainty maps is in* EXPTIME.

PROOF. Given an uncertainty map $\mathcal{M} = \langle \mathcal{N}, \mathcal{U} \rangle$, the construction of ETS $\mathcal{N}^\bullet$ can be done in exponential time in the size of $\mathcal{N}$ due to the fact that there are at most $|\mathcal{N}|$ a-successors t_Δ of each s_Γ since $\Delta = \Gamma|^a$. By modifying the algorithm for PDL in [22], we can get an algorithm to check EPDL formula ϕ on $\mathcal{N}^\bullet$ w.r.t. $\Vdash$, and its time complexity is $O(|\phi|^2 \cdot |\mathcal{N}^\bullet|^3)$. Thus, the time complexity of model checking ϕ on $\mathcal{M}$ is bounded by $O(|\phi|^2 \cdot |\mathcal{S}_\mathcal{N}|^3 \cdot 2^{3|\mathcal{S}_\mathcal{N}|-3})$. □

We conjecture that the model checking problem of full EPDL is EXPTIME-complete, and leave the lower bound to the extended version of this paper.

4.3 Complexity of conformant planning

In the rest of this section, let us look at the complexity of conformant planning in terms of EPDL model checking. Although the model checking problem of full EPDL is likely to be EXPTIME-complete, the complexity of model checking the EPDL formula which encodes the conformant planning problem (cf. Theorem 3.1) is in PSPACE if the goal formula is program-free. More precisely, we can show the following:

THEOREM 4.5. *The problem of model checking* EPDL *formulas in the shape of* $\langle(\Sigma_{a\in B}(?K\langle a\rangle\top; a))^*\rangle K\phi$, *where* ϕ *is an epistemic formula (i.e. program-free) and* B $\subseteq$ A, *is in* PSPACE.

PROOF. (Sketch) Note that $(\sum_{a\in \mathtt{B}}(?K\langle a\rangle\top; a))^*$ is a special program which has only simple epistemic tests depending on the structure of the underlying Kripke model. Now given a Kripke model $\mathcal{N}$ and a set B $\subseteq$ A we can define an ETS model $\mathcal{N}^\circ$ similar to $\mathcal{N}^\bullet$ but with a different definition for the action relations:

$$\mathcal{R}^\circ_a = \{(s_\Gamma, t_\Delta) \mid s \xrightarrow{a} t, \Delta = \Gamma|^a, \forall u \in \Gamma \exists v \ st.\ u \xrightarrow{a} v.\}$$

Note that the extra condition guarantees that the action a is always executable w.r.t. the whole Γ, thus fulfilling the test $?K\langle a\rangle\top$. Now we can have an analog of Corollary 4.2, and reduce the problem of checking $\langle \mathcal{N}, \mathcal{U}\rangle, s \vDash (\sum_{a\in\mathtt{B}}(?K\langle a\rangle\top; a))^* K\phi$ to the reachability problem in $\mathcal{N}^\circ$: whether there is a path from $s_\mathcal{U}$ in $\mathcal{N}^\circ$ such that it can reach a state $t_{\mathcal{U}'}$ where $K\phi$ holds. Since ϕ is $[\pi]$-free, we can check it easily given $\mathcal{U}'$ using polynomial space, thus the main task is to find the reachable $t_{\mathcal{U}'}$. Note that, in the size of $\mathcal{N}$, there are exponentially many such $t_{\mathcal{U}'}$ and the maximal length of the plan is also exponential. However, we do not need to build the whole $\mathcal{N}^\circ$ and the bisection-like algorithm behind the proof of Savitch's Theorem will do the job.[8] More precisely, we first pick up a $t_{\mathcal{U}'}$, and then run the recursive bisection method to see whether $t_{\mathcal{U}'}$ is reachable from $s_\mathcal{U}$ within $2^{|\mathcal{N}|}$ steps. The depth of the recursion is bounded by $log_2(2^{|\mathcal{N}|}) = |\mathcal{N}|$ and at each recursion we need to record the choice of the state which can be encoded by a $(0,1)$-vector using $log_2(2^{|\mathcal{N}|}) = |\mathcal{N}|$ space (plus one bit to record the result). Moreover, at the bottom of the recursion we only need to verify one step reachability, i.e., whether two states in $\mathcal{N}^\circ$ are linked by $\mathcal{R}^\circ_a$, without building the whole $\mathcal{N}^\circ$. Thus the whole procedure of model checking can be done using polynomial space. □

As we mentioned in the introduction, the conformant planning problems in the AI literature are usually given by using state variables and actions with preconditions and (conditional) effects, rather than explicit transition systems. The corresponding explicit transition system can be generated by taking all the possible valuations of the state variables as the state space (an exponential blow up), and computing the transitions among the valuations according to the preconditions and the postconditions of the actions. In terms of the size of explicit transition systems, our above result is consistent with the EXPSPACE complexity result in the AI literature for conformant planning with Boolean and modal goals [20, 8]. Actually, the complexity result of Theorem 4.5 can be strengthened to PSPACE-complete based on the corresponding complexity result in the AI literature.

However, not all the transition systems can be generated in this way since the preconditions and postconditions are (usually) purely propositional and thus two states that share the same valuation must have the same executable actions. In an arbitrary transition system, multiple states with the same valuation may have different available actions due to some underlying protocol or other (external) factors not modelled by basic propositions.

5. CONCLUSIONS AND FUTURE WORK

In this work we first introduce the logical language EAL over uncertainty maps and axiomatize it completely. EAL is then extended to EPDL with programs to specify conformant and conditional plans. We show that the conformant planning problems can be reduced to model checking problems of EPDL. Finally we showed that model checking star-free EPDL over uncertainty maps is PSPACE-complete and model checking the full fragment is in EXPTIME. On the other hand, model checking the conformant planning problem is in PSPACE.

Note that our EPDL is a powerful language which can already express conditional plans, e.g. $(?p; a+?\neg p; b); c$. This suggests that we can use the very EPDL language (EPDL$^-$ is enough) to *verify* plans in contingent planning w.r.t. a variant of the semantics which can handle feedbacks during the execution. In fact, observational power about the availability of the actions has been already incorporated in [33], which can be extended to general feedbacks discussed in the literature of contingent planning (cf. e.g., [10]). On the other hand, to check the *existence* of a conditional plan, we are not sure whether EPDL is expressive enough, as subtleties may arise as in the case of conformant planning. We leave the contingent planning to future work.

Another natural extension is to go probabilistic, and reduce the probabilistic planning over MDP to some model checking problem of the probabilistic version of our EPDL. Our ultimate goal is to cast all the standard AI planning problems into one unified logical framework in order to facilitate careful comparison and categorization. We will then see clearly how the form of the goal formula, the constructor of the plan, and the observational ability matter in the theoretical and practical complexity of planning, in line with the research pioneered in [5].

Acknowledgement.

Quan Yu is supported by NSF Grant No.61463044 and Grant No.[2014]7421 from the Joint Fund of the NSF of Guizhou province of China. Yanjun Li thanks the support

[8] A similar algorithm was used to pinpoint complexity of the conformant planning in AI, cf.[20].

from China Scholarship Council. Yanjing Wang acknowledges the support from ROCS of SRF by Education Ministry of China and the NSSF major project 12&ZD119.

6. REFERENCES

[1] Mikkel Birkegaard Andersen, Thomas Bolander, and Martin Holm Jensen. Conditional epistemic planning. In *Logics in Artificial Intelligence*, pages 94–106. Springer, 2012.

[2] Mikkel Birkegaard Andersen, Thomas Bolander, and Martin Holm Jensen. Don't plan for the unexpected: Planning based on plausibility models. *Logique et Analyse*, 1(1), 2014.

[3] Guillaume Aucher. Del-sequents for regression and epistemic planning. *Journal of Applied Non-Classical Logics*, 22(4):337–367, 2012.

[4] Guillaume Aucher and Thomas Bolander. Undecidability in epistemic planning. In *IJCAI*, pages 27–33, 2013.

[5] Christer Bäckström and Peter Jonsson. All pspace-complete planning problems are equal but some are more equal than others. In *SOCS 2011*, 2011.

[6] Alexandru. Baltag and Larry Moss. Logics for epistemic programs. *Synthese*, 139:165–224, March 2004.

[7] Thomas Bolander and M. Birkegaard Andersen. Epistemic planning for single and multi-agent systems. *Journal of Applied Non-Classical Logics*, 21(1):9–34, 2011.

[8] Blai Bonet. Conformant plans and beyond: Principles and complexity. *Artificial Intelligence.*, 174(3-4):245–269, 2010.

[9] Blai Bonet and Hector Geffner. Planning with incomplete information as heuristic search in belief space. In *ICAPS 2000*, pages 52–61, 2000.

[10] Blai Bonet and Hector Geffner. Width and complexity of belief tracking in non-deterministic conformant and contingent planning. In *AAAI 12*, 2012.

[11] Ronen I. Brafman and Jörg Hoffmann. Conformant planning via heuristic forward search: A new approach. In *ICAPS 2004*, pages 355–364, 2004.

[12] Daniel Bryce, Subbarao Kambhampati, and David E. Smith. Planning graph heuristics for belief space search. *Journal of Artificial Intelligence Research*, 26:35–99, 2006.

[13] Alessandro Cimatti and Marco Roveri. Conformant planning via symbolic model checking. *Journal of Artificial Intelligence Research*, 13:305–338, 2000.

[14] Alessandro Cimatti and Marco Roveri. Conformant planning via symbolic model checking. *CoRR*, abs/1106.0252, 2011.

[15] Alessandro Cimatti, Marco Roveri, and Piergiorgio Bertoli. Conformant planning via symbolic model checking and heuristic search. *Artificial Intelligence*, 159(1-2):127–206, 2004.

[16] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about knowledge.* MIT Press, Cambridge, MA, USA, 1995.

[17] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. Knowledge-based programs. *Distributed Computing*, 10(4):199–225, July 1997.

[18] Patrik Haslum and Peter Jonsson. Some results on the complexity of planning with incomplete information. In *ECP 1999*, pages 308–318, 1999.

[19] Tao Jiang and B. Ravikumar. A note on the space complexity of some decision problems for finite automata. *Information Processing Letters*, 40:25–31, October 1991.

[20] Jon Kleinberg and Éva Tardos. *Algorithm Design.* Addison-Wesley, 2005.

[21] Jérôme Lang and Bruno Zanuttini. Knowledge-based programs as plans - the complexity of plan verification. In *ECAI 2012*, pages 504–509, 2012.

[22] Martin Lange. Model checking propositional dynamic logic with all extras. *Journal of Applied Logic*, 4:39–49, 2006.

[23] Benedikt Löwe, Eric Pacuit, and Andreas Witzel. Del planning and some tractable cases. In *LORI 2011*, pages 179–192. Springer, 2011.

[24] Héctor Palacios and Hector Geffner. Compiling uncertainty away: Solving conformant planning problems using a classical planner (sometimes). In *AAAI 2006*, pages 900–905, 2006.

[25] Pere Pardo and Mehrnoosh Sadrzadeh. Strong planning in the logics of communication and change. In *Declarative Agent Languages and Technologies X*, pages 37–56. Springer, 2013.

[26] Rohit. Parikh and R. Ramanujam. Distributed processes and the logic of knowledge. In *Proceedings of Conference on Logic of Programs*, pages 256–268, London, UK, 1985. Springer-Verlag.

[27] Philippe Schnoebelen. The complexity of temporal logic model checking. In Philippe Balbiani, Nobu-Yuki Suzuki, Frank Wolter, and Michael Zakharyaschev, editors, *AiML 2002*, pages 393–436, Toulouse, France, 2003. King's College Publication. Invited paper.

[28] David E. Smith and Daniel S. Weld. Conformant graphplan. In *AAAI 1998*, pages 889–896, 1998.

[29] Larry J Stockmeyer and Albert R Meyer. Word problems requiring exponential time (preliminary report). In *STOC 1973*, pages 1–9. ACM, 1973.

[30] Son Thanh To, Tran Cao Son, and Enrico Pontelli. A new approach to conformant planning using cnf*. In *ICAPS 2010*, pages 169–176, 2010.

[31] Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic epistemic logic.* Springer, 2007.

[32] Yanjing Wang and Qinxiang Cao. On axiomatizations of public announcement logic. *Synthese*, 190:103–134, 2013.

[33] Yanjing Wang and Yanjun Li. Not all those who wander are lost: Dynamic epistemic reasoning in navigation. In *AiML 2012*, pages 559–580, 2012.

[34] Quan Yu, Ximing Wen, and Yongmei Liu. Multi-agent epistemic explanatory diagnosis via reasoning about actions. In *IJCAI 2013*, pages 1183–1190, 2013.

APPENDIX

A. ALGORITHMS FOR EPDL$^-$

DEFINITION A.1 (MATRIX REPRESENTATION). *Let $\mathtt{B}_{n\times m}$ denote a (0,1)-matrix of size $n \times m$. A matrix $\mathtt{B}_{n\times 1}$, or $\mathtt{B}_n$ for short, is called a vector. Given finite uncertainty map*

$\mathcal{M}$, its domain $\mathcal{S}$ can be linearly ordered as $\{s_1, \cdots, s_n\}$. Thus $\mathcal{M}$ can be represented by a set $\{B^a_{n\times n} \mid a \in A\}$ of adjacency matrices for accessibility relation, a vector $B^{\mathcal{U}}_n$ for $\mathcal{U}$ and a set $\{B^p_n \mid p \in P\}$ of vectors for atomic propositions.

Definition A.2. *Given (0,1)-matrices $B'_{n\times k}, B_{k\times m}$, their product $B''_{n\times m}$ is defined as: $B''_{n\times m}[i,j] = 1$ iff there exists $r \leq n$ such that $B'_{n\times k}[i,r] = B_{k\times m}[r,j] = 1$ for all $1 \leq i \leq n, 1 \leq j \leq m$.*

The following algorithms are to check whether ϕ holds on a pointed uncertainty map $\mathcal{M}, s$ by Definition 4.3. The main algorithm (Algorithm 3) recursively calls itself for each non-trivial subformula of ϕ. The complex cases are for the subformulas in the form of $\langle\pi\rangle\phi$ and $K\phi$. By Definition 4.3, to check $\mathcal{M}, s \Vdash_\sigma \langle\pi\rangle\phi$, we need to make sure that there exists a sequence $\omega \in \mathcal{L}(\pi)$ and a state $t \in \mathcal{S}$ such that $s \overset{\omega_\sigma}{\to} t$ and $\mathcal{M}, t \Vdash_{\sigma r(\omega)} \phi$. Since π is star-free, $|\omega| \leq |\pi|$ for each $\omega \in \mathcal{L}(\pi)$. It is clear that we cannot compute and store the whole set of $\mathcal{L}(\pi)$ within polynomial space. Instead, *one by one* we generate all the possible sequences that are shorter than $|\pi|$ and are formed from the alphabet of π (cf. line 14), and check whether they are in $\mathcal{L}(\pi)$. We can order the possible sequences lexicographically according to an ordering of the basic actions and tests in *Sig*, and compute the next sequence merely from the current one using function *next*. *memb_chec(ω, π)* checks whether it is the case $\omega \in \mathcal{L}(\pi)$. If $\omega \in \mathcal{L}(\pi)$, we need to check whether there exists $s_j \in \mathcal{S}_\mathcal{M}$ such that $s \overset{\omega_\sigma}{\to} s_j$ (Algorithm 2) and $\mathcal{M}, s_j \Vdash_{\sigma r(\omega)} \phi$, where $r(\omega)$ is the test-free subsequence of ω which is easy to compute. For the case of $K\phi$, we need to calculate the state set $\mathcal{U}|^\sigma$ (Algorithm 1).

B. COMPLEXITY ANALYSIS

We suppose $|\mathcal{S}_\mathcal{M}| = n$ and $|\phi| = k$. Algorithm 1 uses one variable A to record the uncertainty set which requires $O(n)$ space. Note that there is a mutual recursion in Algorithm 2 and 3, but the depth of the overall recursion is bounded by k. In Algorithm 2, the variable consuming the most of the space is the matrix $B_{n\times n}$ recording the (intermediate) relation. Since σ and ω are also variables in the main algorithm and $|\omega| + |\sigma| \leq k$ due to the construction in Algorithm 3, the space usage of Algorithm 2 before the recursive calls of *PW* and *MC* is bounded by $O(k+n^2)$. For Algorithm 3, the most space-demanding part is the $\langle\pi\rangle\phi$ case, where we need to store π, *Sig*, and keep track one ω and one state s in the loop, which are bounded by either k or s. Moreover, according to [19], the complexity of *memb_chec* is NLOGSPACE-complete in the size of *Sig*, i.e., the alphabet of π which is bounded again by k. Thus before calling *MC* and *PW* again in the $\langle\pi\rangle\phi$ case, the space requirement is at most linear in both k and n, which is less demanding than *PW* for each recursion. Recall that the overall recursion depth of *MC* (and *PW*) is bounded by k thus the space usage of the whole algorithm is bounded by $O(k(k+n^2)) = O(k^2 + kn^2)$.

Algorithm 1: Function CNU($\mathcal{U}, \sigma$): Calculate the the new uncertainty set $\mathcal{U}|^\sigma$

```
input  : $\mathcal{U}$, $\sigma$
output: $B^{\mathcal{U}|^\sigma}_n$
1 $A \leftarrow B^{\mathcal{U}}_n$;
2 for $i \leftarrow 1$ to $|\sigma|$ do
3     $A \leftarrow A \times B^{\sigma[i]}_{n\times n}$;
4 return $A$;
```

Algorithm 2: Function $PW(\omega, \sigma)$: Calculate the binary relation $\overset{\omega_\sigma}{\to}$

```
input  : computation sequence $\omega$, action sequence $\sigma$
output: $B_{n\times n}$
1 switch $\omega_\sigma$ do
2     case $\epsilon_\sigma$ return $Matrix(\{(s,s) \mid s \in \mathcal{S}\})$
      /* $Matrix(R)$ is the $(0,1)$-matrix
      representation of the binary relation $R$ */;
3     case $(?\phi\omega')_\sigma$
4         return $Matrix(\{(s,s) \mid$ MC$(\mathcal{M}, s, \sigma, \phi) =$ true
          $\}) \times$ PW$(\omega', \sigma)$;
5     case $(a\omega')_\sigma$ return $B^a_{n\times n} \times$ PW$(\omega', \sigma a)$ ;
```

Algorithm 3: Function MC($\mathcal{M}, s, \sigma, \phi$): Model checking algorithm for $\mathtt{EPDL}^-$ (Boolean cases omitted)

```
input  : The pointed uncertainty map $(\mathcal{M}, s)$, sequence
         of actions $\sigma$, $\phi \in \mathtt{EPDL}^-$.
output: true if $\mathcal{M}, s \Vdash_\sigma \phi$.
1  switch $\phi$ do
2      case $\langle\pi\rangle\varphi$
3          Let $Sig$ be the array consisting of atomic
           programs and formulas in $\pi$ ordered according
           to their first appearances;
4          $\omega \leftarrow Sig[1]$ /* $\omega$ is the candidate sequence
           we want to test */;
5          while $|\omega| \leq |\pi|$ do
6              if $memb\_chec(\omega, \pi)$ then
7                  for $i$ = $1$ to $\mathcal{S}_\mathcal{M}$ do
8                      if $(s, s_i) \in PW(\omega, \sigma)$ then
9                          if $MC(\mathcal{M}, s_j, \sigma r(\omega), \varphi)$ then
10                             return true;
11             $\omega \leftarrow next(\omega, Sig)$ /* calculate the next
               sequence lexicographically according
               to the order $Sig$ */;
12         return false;
13     case $K\varphi$
14         $B^{\mathcal{U}|^\sigma}_n$ = CNU$(\mathcal{U}, \sigma)$ /* calculate the vector
           representation of $\mathcal{U}|^\sigma$ */
15         for $m$ = $1$ to $|\mathcal{S}_\mathcal{M}|$ do
16             if $(B^{\mathcal{U}|^\sigma}_n)_m = 1$ and $MC(\mathcal{M}, s_m, \sigma, \varphi) =$
               $false$ then
17                 return false;
18         return true;
```

Distinguishing Cause from Effect Based on Exogeneity

[Extended Abstract]

Kun Zhang
MPI for Intelligent Systems
72076 Tübingen, Germany &
Info. Sci. Inst., USC
4676 Admiralty Way, CA 90292
kzhang@tuebingen.mpg.de

Jiji Zhang
Department of Philosophy
Lingnan University
Hong Kong S.A.R., China
jijizhang@ln.edu.hk

Bernhard Schölkopf
Max-Planck Institute for
Intelligent Systems
72076 Tübingen, Germany
bs@tuebingen.mpg.de

ABSTRACT

Recent developments in structural equation modeling have produced several methods that can usually distinguish cause from effect in the two-variable case. For that purpose, however, one has to impose substantial structural constraints or smoothness assumptions on the functional causal models. In this paper, we consider the problem of determining the causal direction from a related but different point of view, and propose a new framework for causal direction determination. We show that it is possible to perform causal inference based on the condition that the cause is "exogenous" for the parameters involved in the generating process from the cause to the effect. In this way, we avoid the structural constraints required by the SEM-based approaches. In particular, we exploit nonparametric methods to estimate marginal and conditional distributions, and propose a bootstrap-based approach to test for the exogeneity condition; the testing results indicate the causal direction between two variables. The proposed method is validated on both synthetic and real data.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Miscellaneous*

General Terms

Algorithms, Theory

Keywords

Causal discovery, causal direction, exogeneity, statistical independence, bootstrap

1. INTRODUCTION

Understanding causal relations allows us to predict the effect of changes in a system and control the behavior of the system. Since randomized experiments are usually expensive and often impracticable, causal discovery from non-experimental data has attracted much interest [18, 23]. To do this, it is crucial to find (statistical) properties in the non-experimental data that give clues about causal relations. For instance, under the causal Markov condition and faithfulness assumption, the causal structure can be partially estimated by constraint-based methods, which make use of conditional independence relationships.

Here we are concerned with the two-variable case, in which constraint-based methods, such as the PC algorithm [23], do not apply. We assume that the given observations are i.i.d., i.e., there is no temporal information. Recently, causal discovery based on structural equation models (SEMs) has proved useful in distinguishing cause from effect [21, 8, 27, 26, 15, 29]; however, the performance of such approaches depends on assumptions on the functional model class and/or on the data-generating functions. On the other hand, there have been attempts in different fields to characterize properties related to causal systems. One such concept (or family of concepts) is known as *exogeneity*, which is salient in econometrics [3, 4]. Roughly speaking, the notion expresses the property that the process that determines one variable X is in some sense separate from or independent of the process that determines another variable, say Y, given the value of X.

The sense of "separateness" or "independence" in the rough idea has been specified in several ways for different purposes, which result in different concepts of exogeneity. The concept that is most relevant in this paper is the one in the context of model reduction, which was originally proposed as a condition that justifies inferences about the parameters of interest based on the conditional likelihood function rather than the joint likelihood function [7]. Here is the basic idea. Suppose the joint distribution of (X, Y) can be factorized as

$$p(X, Y|\theta, \psi) = p(Y|X, \psi)p(X|\theta). \qquad (1)$$

where the conditional distribution $p(Y|X)$ is parameterized by ψ alone, and the marginal distribution $p(X)$ by θ alone. According to [3, 19], X is said to be exogenous for ψ (or any parameter of interest that is a function of ψ), if ψ and θ are variation free[1], or in other words, are not subject to 'cross-restrictions". From the frequentist point of view, this implies that ψ and θ are independently estimable: the MLE of ψ and that of θ are statistically independent according to the sampling distribution. From the Bayesian point of view [4], this implies that ψ and θ are a posteriori independent given independent priors on them.

In this paper we will exploit the above idea to develop a test of whether *there exists* a parameterization (θ, ψ) for $p(X, Y)$ such that X is exogenous for ψ, the parameters for

[1] This is actually the definition of "weak exogeneity" in [3], where three types of exogeneity were defined. Here we consider the i.i.d. case where there is no temporal information, and consequently strong exogeneity in [3] and weak exogeneity conincide.

$p(Y|X)$. The test is based on bootstrap and is applicable in nonparametric settings. We will also argue that if X is a cause of Y and there is no confounding, then there should exist a parameterization such that X is exogenous for the parameters for $p(Y|X)$. Thus the nonparametric test can be used to indicate the causal direction between two variables, when the test passes for one direction but fails for the other. Compared to the SEM-based approach, an important novelty of this work is to use exogeneity as a new criterion for causal discovery in general settings, which allows distinguishing cause from effect and detecting confounders without structural constraints on the causal mechanism.[2]

2. EXOGENEITY AND CAUSALITY

In this section we define what "exogeneity" means in this paper, and explain its link to causal asymmetry. The concept of exogeneity we will use is adapted from the concept known in econometrics as *weak exogeneity*, which is in itself a statistical rather than a causal concept.[3] We will show that this statistical notion can nonetheless be exploited to formulate a method that can often determine the causal direction between two variables.

2.1 Exogeneity

The concept of weak exogeneity, as formulated by Engle, Hendry, and Richard (EHR) [3], is concerned with when efficient estimation of a set of parameters of interest can be made in a *conditional* submodel. For the purpose of this paper, suppose we are given two continuous random variables X and Y, on which we have i.i.d. observations that are drawn according to a joint density $p(X, Y|\phi)$. By a reparameterization we mean a one-to-one transformation of the parameter set ϕ. Our definition below is adapted from the EHR definition, adjusted for our present purpose and setup:

DEFINITION 1 (EXOGENEITY OF X FOR $p(Y|X)$). *Suppose $p(X, Y)$ is parameterized by ϕ. X is said to be exogenous for the conditional $P(Y|X)$ (or simply, exogenous relative to Y) if and only if there exists a reparameterization $\phi \to (\theta, \psi)$, such that*
(i.) $p(X, Y|\theta, \psi) = p(Y|X, \psi)p(X|\theta)$, and
(ii.) θ and ψ are variation free, i.e., $(\theta, \psi) \in \Theta \times \Psi$, where Θ and Ψ denote the set of admissible values of θ and ψ, respectively.

Here "variation free" means that the possible values that one parameter set can take do not depend on the values of the other set. Clauses *(i.)* and *(ii.)* in Definition 1 are the defining conditions for the concept of a *(classical) cut*: $[(Y|X; \psi), (X; \theta)]$ is said to operate a (classical) cut on $p(X, Y|\theta, \psi)$ if *(i.)* and *(ii.)* are satisfied. The cut implies that the maximum likelihood estimates of θ and ψ can be computed from $p(X|\theta)$ and $p(Y|X, \psi)$, respectively, and so the MLEs $\hat{\theta}$ and $\hat{\psi}$ are independent according to the sampling distribution. The concept of exogeneity formalizes the idea that the mechanism generating the exogenous variable X does not contain any relevant information about the parameter set ψ for the conditional model $p(Y|X)$.

The concept of cut also has a Bayesian version: [4, 16].

DEFINITION 2 (BAYESIAN CUT). $[(Y|X; \psi), (X; \theta)]$ *operates a Bayesian cut on $p(X, Y|\theta, \psi)$ if*
(i.) ψ and θ are independent a priori, i.e., $\psi \perp\!\!\!\perp \theta$,
(ii.) θ is sufficient for the marginal process of generating X, i.e., $\psi \perp\!\!\!\perp X|\theta$, and
(iii.) ψ is sufficient for the conditional process of generating Y given X, i.e., $\theta \perp\!\!\!\perp Y|(\psi, X)$.

A Bayesian cut allows a complete separation of inference (on θ) in the marginal model and of inference (on ψ) in the conditional model. The prior independence between θ and ψ in the Bayesian cut is a counterpart to the variation-free condition in the classical cut, and the last two conditions in Definition 2 implies condition *(i.)* in Definition 1. Thus, the Bayesian cut is equivalent to the classical cut in sampling theory, and for the purpose of this paper can be regarded as interchangeable. Therefore, the exogeneity of X relative to Y can also be defined as that there exists a reparameterization (θ, ψ) of $p(X, Y)$ such that $[(Y|X; \psi), (X; \theta)]$ operates a Bayesian cut on $p(X, Y|\theta, \psi)$.

2.2 Possible Situations Where the Parameterization Fails to Operate a Bayesian Cut

Fig. 1(a) shows a data-generating process of X and Y from where $[(Y|X; \psi), (X; \theta)]$ operates a Bayesian cut. Note that in Definition 2, the two requirements of sufficiency of ψ and θ for the marginal and the conditional (conditions *(ii.)* and *(iii.)*), respectively, are only restrictive under the assumption of prior independence of θ and ψ (condition *(i.)*); otherwise, conditions *(ii.)* and *(iii.)* can be trivially met by, for example, taking θ and ψ to be the same. In fact, any two conditions in Definition 2 could be trivial, given that the other does not hold. Fig. 1(b–d) shows the situations where conditions *(i.)*, *(ii.)*, and *(iii.)* are violated, respectively. In all those situations, θ and ψ are not independent *a posteriori*.

2.3 Relation to Causality

As Pearl [18] rightly stressed, the EHR concept of weak exogeneity is a statistical rather than a causal notion. Unlike the concept of super exogeneity, it is not defined in terms of interventions or multiple regimes. That is why, as we will show, the hypothesis that X is exogenous relative to Y in the sense we defined is generally testable by observational data. However, it is also linked to causality in that it is arguably a necessary condition for an unconfounded causal relation: if X is a cause of Y and there is no common cause of X and Y, then X is exogenous relative to Y in the sense we defined.[4] This follows from the principle we indicated at the beginning: if X is an unconfounded cause of Y, then the process or mechanism that determines X is separate or independent from the process or mechanism that determines Y given X. The separation of processes ensures the *existence* of separate parameterizations of the processes, which will then satisfy our definition of exogeneity.

[2] A related criterion is that of *algorithmic independence* between the input distribution $p(X)$ and the conditional $p(Y|X)$ postulated for a causal system $X \to Y$ [11]; see also [10]. The algorithmic independence condition is defined in terms of Kolmogorov complexity, which is uncomputable, and the method proposed in this paper provides an alternative way to assess the "independence" between $p(X)$ and $p(Y|X)$.

[3] The stronger, causal concept of exogeneity is known as *super exogeneity*.

[4] In this paper we use "unconfounded" to mean the absence of any common cause.

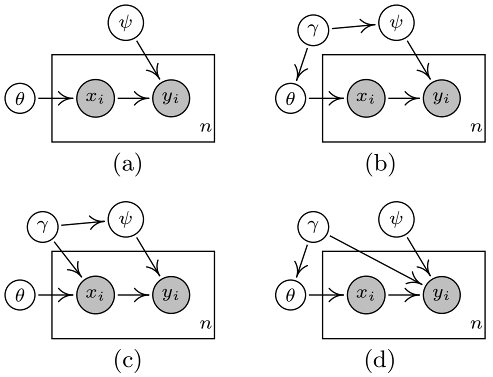


Figure 1: Graphical representation of the data-generating process. (a) $[(Y|X;\psi),(X;\theta)]$ operates a Bayesian cut (implying that X and ψ are mutually exogenous). (b), (c), and (d) correspond to three situations where $[(Y|X;\psi),(X;\theta)]$ does not operate a Bayesian cut: (b) ψ and θ are dependent *a priori*, as both of them depend on γ, which is a function of θ or ψ; (c) θ is not sufficient in modeling the marginal distribution of X, where γ is a function of ψ; (d) ψ is not sufficient in modeling the conditional distribution of Y given X, where γ is a function of θ.

We have argued that if X and Y are causally related and unconfounded, the exogeneity property holds for the correct causal direction. Furthermore, if it turns out that there is one and only one direction that admits exogeneity, then the direction for which the exogeneity property holds must be the correct causal direction. This suggests the following approach to inferring the causal direction between X and Y based on some tests of exogeneity, assuming that X and Y are causally related and that there is no common cause of X and Y (or in other words, X and Y form a *causally sufficient* system): test whether (1) X is exogenous for $p(Y|X)$ and whether (2) Y is exogenous for $p(X|Y)$, and if one of them holds and the other does not, we can infer the causal direction accordingly. Of course it may also turn out that neither (1) nor (2) holds, which will indicate that the assumption of causal sufficiency is not appropriate, or that both (1) and (2) hold, which will indicate that the causal direction in question is not identifiable by our criterion.[5]

A familiar example of a non-identifiable situation is when X and Y follow a bivariate normal distribution. In that case, as shown by EHR [3], there is a cut $[(Y|X;\psi),(X;\theta)]$ in one direction, as well as a cut $[(X|Y;\tilde{\psi}),(Y;\tilde{\theta})]$ in the other. Below we give an example where the causal direction is identifiable based on exogeneity.

An example of identifiable situation: Linear non-Gaussian case.

Let X follow a Gaussian mixture model with two Gaussians, $X \sim \sum_{i=1}^{2} \pi_i \mathcal{N}(u_i, \sigma_i^2)$, where $\pi_i > 0$ and $\pi_1 + \pi_2 = 1$, and let $Y = c + \beta X + E$ where $E \sim \mathcal{N}(0, \sigma^2)$. Therefore

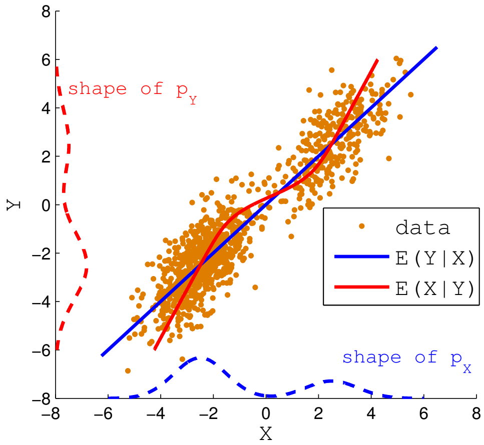


Figure 2: An illutration on the identifiability of a linear non-Gaussian model based on "exogeneity". X is generated by a mixture of two Gaussians, and Y is generated by $Y = X + E$, where $E \sim \mathcal{N}(0,1)$. Here X is exogenous for parameters in $p_{Y|X}$, while Y is not exogenous for parameters in $p_{X|Y}$.

$\theta = \{\pi_i, \mu_i, \sigma_i\}_{i=1}^{2}$ and $\psi = \{c, \beta, \sigma^2\}$. We then have

$$\begin{aligned} p(X,Y|\theta,\beta) &= \sum_i \pi_i \mathcal{N}(x;\mu_i,\sigma_i^2)\mathcal{N}(y;c+\beta x,\sigma^2) \\ &= \sum_i \pi_i \mathcal{N}(y;\tilde{\mu}_i,\tilde{\sigma}_i^2)\mathcal{N}(x;\tilde{c}_i+\tilde{\beta}_i y,\gamma_i^2), \end{aligned}$$

where $\tilde{\mu}_i = c + \beta\mu_i$, $\tilde{\sigma}_i^2 = \beta^2\sigma_i^2 + \sigma^2$, $\tilde{c}_i = \frac{\mu_i\sigma^2 - c\beta\sigma_i^2}{\beta\sigma_i^2+\sigma^2}$, $\tilde{\beta}_i = \frac{\beta\sigma_i^2}{\beta\sigma_i^2+\sigma^2}$, and $\gamma_i^2 = \frac{\sigma^2\sigma_i^2}{\beta\sigma_i^2+\sigma^2}$. That is,

$$Y \sim \sum_{i=1}^{2} \pi_i \mathcal{N}(y;\tilde{u}_i,\tilde{\sigma}_i^2), \quad \text{and}$$

$$p(X|Y,\theta,\psi) = \sum_i \frac{\pi_i \mathcal{N}(y;\tilde{\mu}_i,\tilde{\sigma}_i^2)}{p(Y|\theta,\psi)} \cdot \mathcal{N}(x;\tilde{c}_i+\tilde{\beta}_i y,\gamma_i^2).$$

Clearly, if $\pi_1\pi_2 \neq 0$, no matter how one parametrizes the density of Y, the conditional distribution of X given Y would involves those parameters that model the marginal density of Y. The sufficient parameter set of the distribution of Y, $\tilde{\theta}$, and that of the conditional distribution of X given Y, $\tilde{\psi}$, cannot be variation-free or independent *a priori*; see Fig. 1(b). Alternatively, one can keep those parameters that are independent *a priori* from $\tilde{\theta}$ in $\tilde{\psi}$, i.e., $\tilde{\psi}$ and $\tilde{\theta}$ become independent *a priori*, but $\tilde{\psi}$ is then not sufficient in modeling $p(X|Y)$; see Fig. 1(d). In both situations Y is not exogenous for $\tilde{\psi}$. Hence in this linear non-Gaussian case the exogeneity condition only holds for the direction $X \rightarrow Y$, and the causal direction is identifiable. Fig. 2 gives an intuitive illustration on how the shape of $P(Y)$ and that of $E(X|Y)$, which is determined by $P(X|Y)$, are related.

3. CAUSAL DIRECTION DETERMINATION BY TESTING FOR EXOGENEITY WITH BOOTSTRAP

[5] Note that we are not concerned with the case in which X and Y are not causally connected and hence statistically independent; in that case, exogeneity trivially holds in both directions.

We now describe our approach to testing exogeneity. We will first illustrate how bootstrap can be used to test whether a given parametric model constitutes a (Bayesian) cut, and then develop a nonparametric test for exogeneity based on bootstrap.

3.1 Bootstrap-Based Test for Bayesian Cut in the Parametric Case

In this section, we assume that a parametric form $p(X, Y|\theta, \psi) = p(X|\theta)p(Y|X, \psi)$ is given. We would like to see whether the estimates of θ and of ψ in (1) are independent, according to the sampling distribution; in other words, with a noninformative prior, we want to test if the posterior distribution $p(\theta, \psi|\mathcal{D})$ has no coupling between θ and ψ. In this case we are examining if $[(Y|X; \psi), (X; \theta)]$ operates a Bayesian cut.

Bootstrap has been used in the literature to assess the dependence, as well as uncertainty, in the parameter estimates according to the sampling distribution; see e.g. [2, Sec. 5.7]. For clarity, Table 1 gives the notation used in the proposed bootstrap-based method. Suppose we draw bootstrap resamples $(\mathbf{x}^{*(b)}, \mathbf{y}^{*(b)})$, $b = 1, ..., B$, from the original sample $(\mathbf{x}, \mathbf{y}) = (x_i, y_i)_{i=1}^N$ with paired bootstrap, i.e., each resample $(\mathbf{x}^{*(b)}, \mathbf{y}^{*(b)})$ is obtained by independently drawing N pairs from the original sample with replacement. On each of them, we can calculate the parameter estimates $\hat{\theta}^{*(b)}$ and $\hat{\psi}^{*(b)}$. The independence between θ and ψ according to the sampling distribution is then transformed to statistical independence between the bootstrap estimates $\hat{\theta}^{*(b)}$ and $\hat{\psi}^{*(b)}$, $b = 1, ..., B$. To assess the latter, any independence test method, such as the correlation test, would apply.

3.2 Bootstrap-Based Test for Exogeneity in the Nonparametric Case

Let $\tilde{\mathbf{x}}$ be a fixed set of values of X, and $\tilde{x}_i$ be a point in $\tilde{\mathbf{x}}$. $\tilde{\mathbf{x}}$ can be drawn from the given data set, or randomly sampled on the support of X, given that it contains enough points such that the values of $P(X)$ and $p(Y|X)$ evaluated at $\tilde{\mathbf{x}}$ well approximate the continuous densities. In our experiments we used 80 evenly-spaced sample points between the minimum and maximum values of X as $\tilde{\mathbf{x}}$ (so its length is $N = 80$).

On the bootstrap resamples, $\log \hat{p}^{*(b)}(X = \tilde{\mathbf{x}})$ is fully determined by $\hat{\theta}^{*(b)}$; similarly, $\log \hat{p}^{*(b)}(Y|X = \tilde{\mathbf{x}})$ is a function of $\hat{\psi}^{*(b)}$, and so is the quantity $H_{Y|X}^{*(b)}(\tilde{\mathbf{x}}) \triangleq \mathbb{E}_{Y|X} \log \hat{p}^{*(b)}(Y|X = \tilde{\mathbf{x}})$. Note that $\hat{p}^{*(b)}(Y|X = \tilde{x}_i)$ is the estimated distribution of Y at $X = \tilde{x}_i$, and hence $H_{Y|X}^{*(b)}(\tilde{\mathbf{x}})$ can be considered as negative entropies of Y on the bth bootstrap resample evaluated at $X = \tilde{\mathbf{x}}$.

Suppose all involved parameters are identifiable, i.e., the mappings $\theta \mapsto p(X|\theta)$ and $\psi \mapsto p(Y|X, \psi)$ are both one-to-one [14]. Then the mapping between $\hat{\theta}^{*(b)}$ and $\log \hat{p}^{*(b)}(X = \tilde{\mathbf{x}})$ and that between $\hat{\psi}^{*(b)}$ and $\log \hat{p}^{*(b)}(Y|X = \tilde{\mathbf{x}})$ are both one-to-one. Hence, the independence between $\hat{\theta}^{*(b)}$ and $\hat{\psi}^{*(b)}$, $b = 1, ..., B$, implies that between $\log \hat{p}^{*(b)}(X = \tilde{\mathbf{x}})$ and $H_{Y|X}^{*(b)}(\tilde{\mathbf{x}})$.

As a consequence, in nonparametric settings, we can imagine that there exist effective parameters θ and ψ, and can still assess where they follow a Bayesian cut by testing for independence between the bootstrapped estimates $\log \hat{p}^{*(b)}(X = \tilde{\mathbf{x}})$ and $H_{Y|X}^{*(b)}(\tilde{\mathbf{x}})$. Note that in the nonparametric case, the "parameters" θ and ψ are not observable. The previous argument shows that if there exists (θ, ψ) admitting a Bayesian cut, $\log \hat{p}^{*(b)}(X = \tilde{\mathbf{x}})$ and $H_{Y|X}^{*(b)}(\tilde{\mathbf{x}})$ are independent; otherwise they are always dependent. In words, testing for independence between the bootstrapped estimates $\log \hat{p}^{*(b)}(X = \tilde{\mathbf{x}})$ and $H_{Y|X}^{*(b)}(\tilde{\mathbf{x}})$ is actually a ways to assess the exogeneity condition. Algorithm 1 sunmmarizes the proposed procedure to determine the causal direction between X and Y, given the sample $(\mathbf{x}, \mathbf{y})$ as input. In particular, it involves the following two modules.

Algorithm 1 Finding causal direction between X and Y based on exogeneity

```
Input: data $(\mathbf{x}, \mathbf{y})$
Output: three possibilities: causal direction between $X$ and $Y$, or non-identifiable causal direction by exogeneity, or existence of hidden confunders
IF_EXOGENEITY($X \to Y$)
IF_EXOGENEITY($Y \to X$)
if exogeneity holds for only one direction then
    return the direction in which exogeneity holds
else if exogeneity holds for both directions then
    print non-identifiable causal direction by exogeneity
else        ▷ exogeneity does not hold in either direction
    print confounder case
end if

procedure IF_EXOGENEITY($X \to Y$)
    for $b = 1$ to $B$ do
        draw bootstrap resample $(\mathbf{x}^{*(b)}, \mathbf{y}^{*(b)})$ by random sampling with replacement from $(x_i, y_i)$;
        estimate $\hat{p}_X^{*(b)}(X = \tilde{\mathbf{x}})$ and $H_{Y|X}^{*(b)}(\tilde{\mathbf{x}})$ with methods given in Sec. 3.2.1
    end for
    test for independence between $\hat{p}_X^{*(b)}(X = \tilde{\mathbf{x}})$ and $H_{Y|X}^{*(b)}(\tilde{\mathbf{x}})$, $b = 1, ..., B$, with the method given in Sec. 3.2.2
    return independence test result
end procedure
```

3.2.1 Module 1: Nonparametric Estimators of $p(X)$ and $p(Y|X)$

When testing for exogeneity, one assumes the (parametric) model is correctly specified. Otherwise, if the model is over-simplified, the estimated conditional distribution will depend on the marginal, which inspires the importance-reweighting scheme to handle learning problems under covariate shift (see e.g., Footnote 1 in [24]). For example, let us consider the situation where Y depends on X in a nonlinear manner while a linear model is exploited to estimate $p_{Y|X}$; clearly the estimate of the parameters in the conditional model would depend on that in p_X. To avoid this, we use flexible nonparametric models to estimate the conditional.

Suppose we aim to verify if X exogenous for effective "parameters" in $P(Y|X)$. We need to estimate the marginal distribution $p(X)$ and the conditional distribution $p(Y|X)$ on the original sample as well as each bootstrap resample. We estimate $p(X)$ with Gaussian kernel density estimation, and the kernel width was selected by Silverman's rule of thumb [22, page 48].

To estimate the conditional density $p(Y|X)$, we adapted

Table 1: Notation involved in the proposed method based on exogeneity and bootstrap

$(\mathbf{x}, \mathbf{y})$	given sample of (X, Y)
$(\mathbf{x}^{*(b)}, \mathbf{y}^{*(b)})$	bth bootstrap resample
$\hat{\theta}^{*(b)}, \hat{\psi}^{*(b)}$	estimate of parameters θ and ψ on $(\mathbf{x}^{*(b)}, \mathbf{y}^{*(b)})$
$\hat{p}^{*(b)}(X = \tilde{\mathbf{x}})$	marginal densities estimated on $(\mathbf{x}^{*(b)}, \mathbf{y}^{*(b)})$ evaluated at $X = \tilde{\mathbf{x}}$
$\hat{p}^{*(b)}(Y \vert X = \tilde{\mathbf{x}})$	conditional densities estimated on $(\mathbf{x}^{*(b)}, \mathbf{y}^{*(b)})$ evaluated at $X = \tilde{\mathbf{x}}$
$H^{*(b)}_{Y \vert X}(\tilde{\mathbf{x}})$	quantity associated with $\hat{p}^{*(b)}(Y \vert X = \tilde{\mathbf{x}})$, defined as $\mathbb{E}_{Y \vert X} \log \hat{p}^{*(b)}(Y \vert X = \tilde{\mathbf{x}})$ on $(\mathbf{x}^{*(b)}, \mathbf{y}^{*(b)})$

the method orignally proposed for causal inference based on the structural equation $Y = f(X, E)$ [15]. This method aims to find the functional causal model $Y = f(X, E)$, where $E \perp\!\!\!\perp X$, given $(\mathbf{x}, \mathbf{y})$. Without loss of generality, one can assume that $E \sim \mathcal{N}(0, 1)$. (Otherwise, one can always write $E = g(\tilde{E})$ where g is some appropriate function and $\tilde{E} \sim \mathcal{N}(0, 1)$, and use the functional causal model $Y = f\big(X, g(\tilde{E})\big)$ instead.) Here f is completely nonparametric: it takes a Gaussian process prior with zero mean function and covariance function $k\big((x, e), (x', e')\big)$, where k is a Gaussian kernel, and (x, e) and (x', e') are two points of (X, E). Like in [13], this method optimizes the values of E, denoted by $\hat{e}_i$, as well as involved hyperparameters, and produces the maximum a posterior (MAP) solution of f, by maximizing the approximate marginal likelihood. The functional causal model implies the conditional density:

$$P(Y|X) = \frac{p(X, Y)}{p(X)} = \frac{p(X, E)\big/|\frac{\partial f}{\partial E}|}{p(X)} = p(E)\Big/|\frac{\partial f}{\partial E}|.$$

Finally, once we have the $\hat{e}_i$ and the estimate of f, the conditional density at each point can be estimated as $p(Y = y_i | X = x_i) = p(E = \hat{e}_i)\big/\Big|\frac{\partial f}{\partial E}(x_i, \hat{e}_i)\Big|$.

3.2.2 Module 2: Testing for Independence Between High-Dimensional Vectors

The task is then to test for independence between the estimated quantities on the bootstrap resamples, $\log \hat{p}^{*(b)}(X = \tilde{\mathbf{x}})$ and $H^{*(b)}_{Y|X}(\tilde{\mathbf{x}})$, $b = 1, ..., B$. Their dimentions are the number of data points in $\tilde{\mathbf{x}}$, which is 80 in our experiments.

Let $\mathbf{R}$ be the matrix consisting of the centered version of $\log \hat{p}^{*(b)}(X = \tilde{x}_i)$, obtained on all bootstrap resamples, i.e., the (i, b)th entry of $\mathbf{R}$ is

$$R_{ib} \triangleq \log \hat{p}(X^{*(b)}(X = \tilde{x}_i) - \frac{1}{B} \sum_{k=1}^{B} \log \hat{p}^{*(k)}(X = \tilde{x}_i).$$

Similarly, $\mathbf{S}$ contains the centered version of $H^{*(b)}_{Y|X}(\tilde{x}_i)$, i.e.,

$$S_{ib} \triangleq H^{*(b)}_{Y|X}(\tilde{x}_i) - \frac{1}{B} \sum_{k=1}^{B} H^{*(k)}_{Y|X}(\tilde{x}_i).$$

Both $\mathbf{R}$ and $\mathbf{S}$ are of the size $N \times B$. We define the statistic as $C_{X \to Y} \triangleq \mathrm{Tr}\big((\mathbf{R}\mathbf{S}^T)(\mathbf{R}\mathbf{S}^T)^T\big) = \mathrm{Tr}(\mathbf{R}^T\mathbf{R} \cdot \mathbf{S}^T\mathbf{S})$, which is actually the sum of squares of the covariances between all rows of $\mathbf{R}$ and those of $\mathbf{S}$. The distribution of this statistic under the null hypothesis that $\log \hat{p}^{*(b)}(X = \tilde{\mathbf{x}})$ and $H^{*(b)}_{Y|X}(\tilde{\mathbf{x}})$ are independent can then be constructed by permutation test.

Note that this statistic is actually the Hilbert-Schmidt independence criterion (HSIC) [6] with a linear kernel. That is, we care about linear dependence between $\log \hat{p}^{*(b)}(X = \tilde{\mathbf{x}})$ and $H^{*(b)}_{Y|X}(\tilde{\mathbf{x}})$; this is reasonable because they are in the vicinity of the maximum likelihood estimates and their dependence can be captured by linear approximation. On the other hand, if we use HSIC with Gaussian kernels, the result will be sensitive to the kernel width because the data dimension (the number of rows of $\mathbf{R}$ and $\mathbf{S}$) is high.

4. EXPERIMENTS

In this section we first evaluate the behavior of the proposed bootstrap-based method for causal inference with synthetic data, on which the ground-truth is known, and then apply it on real data. We use two variables, and with synthetic data, we examine both the case where the two variables have a direct causal relation and the confounder case (i.e., there are confounders influencing both of them). We compare the proposed bootstrap-based approach with the additive noise model (ANM) proposed in [8]), GPI [15], and information-geometric causal inference (IGCI) approach [10]: ANM assumes that the effect is a nonlinear function of the cause plus additive noise, GPI applies the Gausian Process prior on the generating function, and IGCI assumes the transformation from the cause to the effect is deterministic, nonlinear, and independent from the distribution of the cause in a certain way. For computational reasons, we used 1000 bootstrap replications.

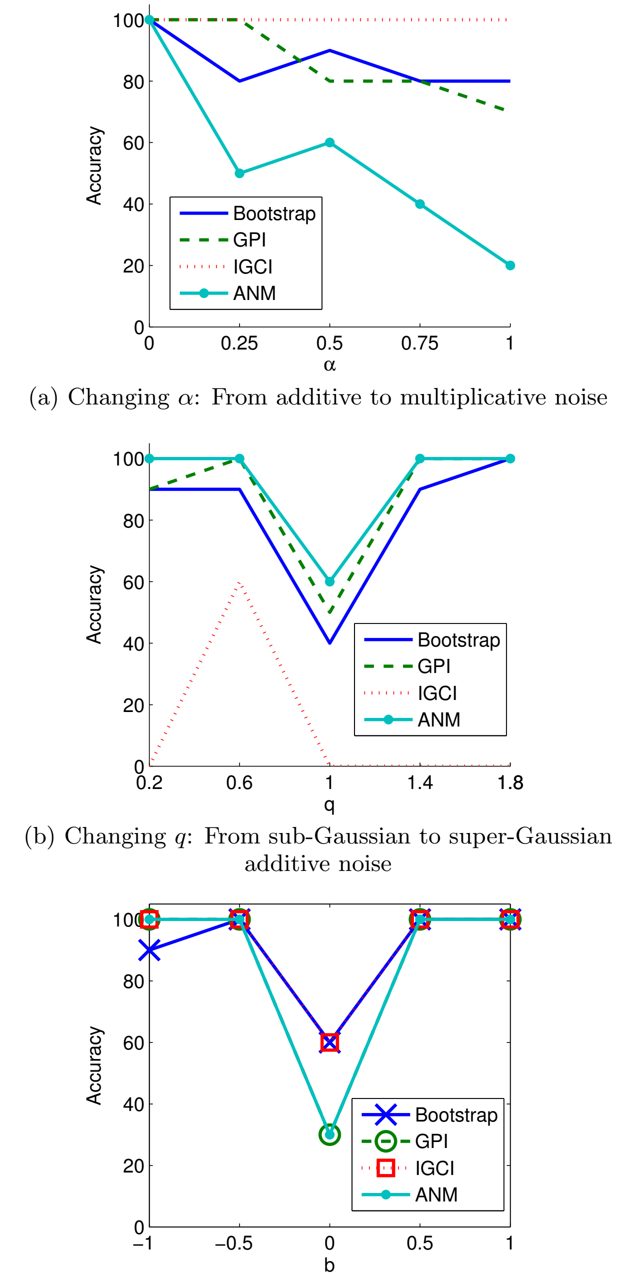
Simulation: Without Confounders. Inspired by the settings in [8, 15], we generated the simulated data with the model $Y = (X + bX^3)e^{\alpha E} + (1 - \alpha)E$, where X and E were obtained by passing i.i.d. Gaussian samples through power nonlinearities with exponent q, while keeping the original signs. The parameter α controls the type of the observation noise, ranging from purely additive noise ($\alpha = 0$) to purely multiplicative noise ($\alpha = 1$). b determines how nonlinear the effect of X is, and when $b = 0$ the model is linear. The parameter q controls the non-Gaussninity of X and E: $q = 1$ corresponds to a Gaussian distribution, and $q > 1$ and $q < 1$ produce super-Gaussian and sub-Gaussian distributions, respectively.

We considered three situations, in each of which two of q, b, and α were fixed and we see how the other changes the performance of different methods. For each combination of q, b, and α, we independently simulated 10 data sets with 500 data points.[6] Fig. 3 shows the accuracy of the considered methods. One can see that the accuracy of the bootstrap-based approach is among or close to the best results, indicating that it is able to perform causal inference in various situations. We note that in practice, the performance of the bootstrap-based approach depends on the number of bootstrap replications and the method used for conditional distribution estimation. Although due to computatioanl reasons, we did not try a larger number of boot-

[6] Since the bootstap-based approach is rather time-consuming, we only simulated 10 data sets for each setting.

strap replications, generally speaking, the accuracy of the bootstrap-based method improves as the number of replications increases.

Simulation: With Confounders. We then include the confounder variable Z in the system, so that the causal structure is $Z \to X$ and $(Z, X) \to Y$. For simplicity, we assume that both X and Y are influenced by Z in a linear form: $X = (2-\beta)E_X + \beta Z$, and $Y = 0.3(2-\beta)\big[(X+bX^3)e^{\alpha E}+(1-\alpha)E\big] + \beta Z$, where E_X, Z, and E were obtained by passing i.i.d. Gaussian samples through power nonlinearities with exponent $q = 1.5$, and β controls how strong the effect of Z is on both X and Y. We considered two situations: in one of them, we set $\alpha = 0$ and $b = 0$, i.e., the whole model is linear; in the other situation, $\alpha = 0.2$, and $b = 0.3$, so the model contains both additive noise and multiplicative noise. We changed β from 0 to 1, and Fig. 4 shows the performances of the four methods in the two situations; note that for each value of β, the four bars (from left to right) correspond to the bootstrap-based method, GPI, IGCI, and ANM. In particular, one can see that the bootstrap-based method tends to detect the presence of the confounder when its effect is significant.

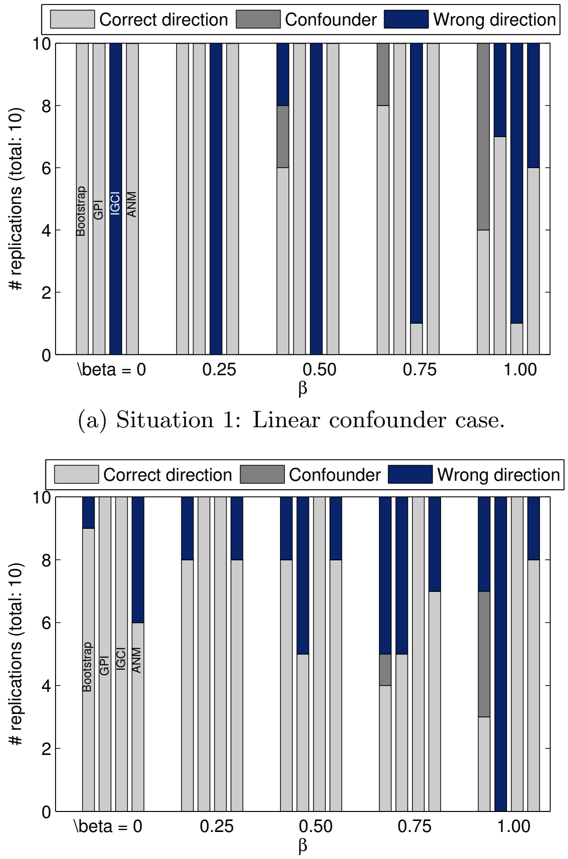


(a) Changing α: From additive to multiplicative noise

(b) Changing q: From sub-Gaussian to super-Gaussian additive noise

(c) Changing b: Various nonlinear functions with Gaussian additive noise

Figure 3: Accuracy of correctly estimating the causal direction for different generating models: (a) $q = 1$, $b = 1$, and α changed from 0 to 1, (b) for a linear function ($b = 0$) with additive noise ($\alpha = 0$) which changed from sub-Gaussian ($q < 1$) to sub-Gaussian ($q > 1$), and (c) various nonlinear functions (b changed from -1 to 1) with additive Gaussian noise ($q = 1$, $\alpha = 0$).



(a) Situation 1: Linear confounder case.

(b) Situation 2: Nonlinear confounder case.

Figure 4: Number of replications in which the methods find correct directions, report existence of confounders, and give wrong directions, respectively. *For each value of β, the four bars correspond to bootstrap-based method, GPI, IGCI, and ANM (from left to right).*

On Real Data. We applied the bootstrap-based method on the cause-effect pairs available at `http://webdav.tuebingen.mpg.de/cause-effect/`. To reduce computational load, we used at most 500 points for each cause-effect pair. On 20 pairs (pairs 21, 43, 45, 48-51, 56-58, 61-64, 72, 75, 77-79, and 81), the p-values of the independence test for both directions are smaller than 0.01, indicating that there might be significant confounders. This seems reasonable, as the data scatter plots for these pairs indicate that the two variables have complex dependence relationships. On the remaining 57 data sets, the bootstrap-based method output correct causal directions on 41 of them (with an accuracy 72%). We also applied the recently proposed causal inference approaches, including IGCI [10], the approach based on the Gaussian process prior [15], and that based on the post-nonlinear causal model [27] on those 57 data sets for comparison. Their performance was similar: the three approaches gave correct causal directions on 41, 40, and 43 pairs, respectively.

5. CONCLUSION AND DISCUSSIONS

We proposed to do causal inference based on the criterion of exogeneity of the cause for the parameters in the conditional distribution of the effect given the cause. We discussed how to assess such exogeneity in nonparametric settings. To this end, one needs to draw a number of samples according to the unknown data-generating process. Fortunately, the bootstrap provides a way to mimic the data generating process from which we can draw a number of samples and analyze their statistical properties.

Our approach shows that it is possible to determine causal direction without structural constraints or a specific type of smoothness assumptions on the functional models. The proposed computational approach successfully demonstrated the validity of this idea, though it is computationally demanding because of the bootstrap procedure and its performance is not necessarily the best among existing methods. At the same time, it enjoys some advantages. First, it does not make a strong assumption on the data-generating process. Second, it could often tell us if significant confounders exist. The performance of the proposed bootstrap-based approach depends on the number of bootstrap replications and the method for conditional distribution estimation. In future work we aim to develop more reliable methods along this line, including methods that can handle more than two variables.

In this paper we made an attempt to discover causal information from observational data based on a condition of exogeneity, which provides another perspective to conceptualize the ”independence” between the process generating the cause and that generating the effect from cause. On the other hand, it is worth mentioning that this type of independence is able to facilitate understanding and solving some machine learning or data analysis problems. For instance, it helps understand when unlabeled data points will help in the semi-supervised learning scenario [20], and inspired new settings and formulations for domain adaptation by characterizing what information to transfer and how to do so [28, 25].

Acknowledgement

We would like to thank Peter Spirtes, Kevin Kelly, and Dominik Janzing for helpful discussions. K. Zhang was supported in part by DARPA grant No. W911NF-12-1-0034. The research of J. Zhang was supported in part by the Research Grants Council of Hong Kong under the General Research Fund LU342213.

6. REFERENCES

[1] W. A. Coppel. *Number theory: An introduction to mathematics.* Springer, second edition, 2009.

[2] B. Efron. *The Bootstrap, Jacknife and other Resampling Plans.* SIAM, Philadelphia, 1982.

[3] R. F. Engle, D. F. Hendry, and J. F. Richard. Exogeneity. *Econometrica*, 51:277–304, 1983.

[4] J. P. Florens and M. Mouchart. Conditioning in dynamic models. *Journal of Time Series Analysis*, 6(1):15–34, 1985.

[5] J. P. Florens, M. Mouchart, and J. M. Rolin. *Elements of Bayesian Statistics.* Marcel Dekker, New York, 1990.

[6] A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In S. Jain, H.-U. Simon, and E. Tomita, editors, *Algorithmic Learning Theory: 16th International Conference*, pages 63–78, Berlin, Germany, 2005. Springer.

[7] D. Henry. *Econometrics: Alchemy or Science?* Oxford University Press, Oxford, 2000.

[8] P. Hoyer, D. Janzing, J. Mooji, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, Vancouver, B.C., Canada, 2009.

[9] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

[10] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniuvsis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, pages 1–31, 2012.

[11] D. Janzing and B. Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56:5168–5194, 2010.

[12] R. Kass, L. Tierney, and J. Kadane. Asymptotics in Bayesian computation. In J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, editors, *Bayesian statistics 3*, pages 261–278. Oxford University Press, 1988.

[13] N. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

[14] E. L. Lehmann and G. Casella. *Theory of point estimation.* Springer, 2nd edition, 1998.

[15] J. Mooij, O. Stegle, D. Janzing, K. Zhang, and B. Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, Curran, NY, USA, 2010.

[16] M. Mouchart and E. Scheihing. Bayesian evaluation of non-admissible conditioning. *Journal of Econometrics*, 123(2):283–306, 2004.

[17] G. H. Orcutt. Toward partial redirection of

econometrics. *The Review of Economics and Statistics*, 34:195–200, 1952.

[18] J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge, 2000.

[19] J. F. Richard. Models with several regimes and changes in exogeneity. *The Review of Economics Studies*, 47:1–20, 1980.

[20] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proc. 29th International Conference on Machine Learning (ICML 2012)*, Edinburgh, Scotland, 2012.

[21] S. Shimizu, P. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

[22] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall/CRC, London, 1998.

[23] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT Press, Cambridge, MA, 2nd edition, 2001.

[24] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.

[25] K. Zhang, , M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 3150–3157. AAAI Press, 2015.

[26] K. Zhang and A. Hyvärinen. Acyclic causality discovery with additive noise: An information-theoretical perspective. In *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 2009*, Bled, Slovenia, 2009.

[27] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, 2009.

[28] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning, JMLR: W&CP Vol. 28*, 2013.

[29] K. Zhang, Z. Wang, J. Zhang, and B. Schölkopf. On estimation of functional causal models: General results and application to post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technologies*, 2015. forthcoming.

Supplement to
"Distinguishing Cause from Effect Based on Exogeneity"

This supplementary material provides the proofs and discussions which are omitted in the submitted paper. The equation numbers in this material are consistent with those in the paper.

S1. Mutual Exogeneity and Its Relationship to Definition 1

There are two types of analysis of exogeneity [4]; one considers the inference based on the complete sample results, and the other considers dynamic models where the data were obtained by "sequential sampling". In this paper we focus on the former scenario.

From the Bayesian point of view, exogeneity of X for ψ allows an admissible reduction of the complete model $p(X, Y|\theta, \psi)$ to the conditional model $p(Y|X, \psi)$, in that both models lead to he same posterior distribution on the parameter set ψ [4, 16]. Below we give the definition of mutual exogneneity according to [4].

Definition 3 (Mutual exogeneity). *X and ψ are mutually exogenous if and only if*

(i) ψ and X are independent, i.e., $\psi \perp\!\!\!\perp X$, and

(ii) ψ is sufficient in the conditional distribution of Y given X, i.e., $\theta \perp\!\!\!\perp Y|(\psi, X)$.

Here condition *(i)* is to do with the independence between ψ and X; those two quantities play different roles in the model $p(X, Y, \psi, \theta)$, and consequently this independence condition is usually not convenient to verify. Moreover, for the same reason, there is no fully equivalent concept in sampling theory (it is weaker than exogeneity defined in Definition 1, because the property of θ is not specified). A natural way of obtaining the mutual exogeneity of X and ψ is to exploit a stronger but more operational condition, namely the condition of the Bayesian cut.

A Bayesian cut allows a complete separation of inference (on parameters θ) in the marginal distribution and of inference (on ψ) in the conditional one. The prior independence between θ and ψ in the Bayesian cut is a counterpart to the variation-free condition in the classical cut (condition *(ii)* in Definition 1), and the last two conditions in Definition 2 implies condition *(i)* in Definition 1. Thus, the Bayesian cut is equivalent to the classical cut in sampling theory, and consequently characterizes the exogeneity property defined in Definition 1. Therefore, hereafter the exogeneity of X for ψ is used interchangeably with the statement that $[\psi, (X, \theta)]$ operates a Bayesian cut in $p(X, Y, \theta, \psi)$.

The following theorem, extracted from [5], relates the Bayesian cut to the independence of the parameters according to the posterior distribution, as well as mutual exogeneity.

Theorem 4. *Suppose $[\psi, (X, \theta)]$ operates a Bayesian cut in $p(X, Y, \{\psi, \theta\})$; then*

(i) X and ψ are mutually exogenous, and

(ii) ψ and θ are independent a posteriori.

On the other hand, if X and ψ are mutually exogenous and if $\theta \perp\!\!\!\perp \psi|X$, $[\psi, (X, \theta)]$ operates a Bayesian cut.

When one (or more) condition in Definition 2 is violated, $[\psi, (X, \theta)]$ does not operate a Baysian cut, i.e., X is not exogenous for ψ. Fig. 1(b–d) shows the situations where conditions *(i)*, *(ii)*, and *(iii)* are violated, respectively, so that $[\psi, (X, \theta)]$ does not operate a Baysian cut. Note that by reparameterization, the three situations can reduce to each other. Take situations (b) and (c) as an example. If we divide θ in (b) into $(\theta_\gamma, \theta_{\perp\!\!\!\perp})$, where θ_γ depends on γ while $\theta_{\perp\!\!\!\perp}$ does not, and consider $\theta_{\perp\!\!\!\perp}$ as the new θ, (b) becomes (c). Similarly, if we merge γ and θ in (c) as the new θ, we then have (b). In all those situations, θ and ψ are not independent *a posterior*, or the maximum likelihood estimates $\hat{\theta}$ and $\hat{\psi}$ are not independent according to the sampling distribution.

S2. Relation to SEM-Based Causal Inference

S2.1. Relation to Causal Inference Based on Marginal Likelihood

Recently, SEM-based approaches have demonstrated their power for causal inference of real-world problems. Structural equations represent the effect as a function of the causes and independent noise, which, from another point of view, provide a way to represent the conditional distribution $P(\texttt{effect}|\texttt{cause})$, or the causal mechanism. The generation of the cause-effect pair consists of two stages, one generating the cause according to $P(\texttt{cause})$ and the other further generating the effect from the value of the cause according to the structural equation. The "simplicity" constraints (e.g., linearity in [21], additive noise in [8], the post-nonlinear process in [27], and the smoothness assumption in [15]) on the functions are crucial. On the one hand, they make the models asymmetric in cause and effect; otherwise, for any two variables, we can always represent one of the variables as a function of the other and an independent noise term [9]. On the other hand, if the functions are constrained to be simple, the independence between the cause and the error terms would imply the exogeneity of the cause for the parameters in $P(\texttt{cause})$, as suggested by the error-based definition of exogeneity [17] (see also [18]).[7]

[7] An error-based definition of exogeneity was given by [17] (see also [18]): X is said to be exogenous for parameters in $p(Y|X)$ is X is independent of all errors that influence Y, except those mediated by X. We know that without appropriate constraints on the functions, given any two random variable, we can always represent one of them as a function of the other variable and an independent noise term [9], i.e., the functional causal models are not identifiable. Therefore,

The concept "exogeneity" provides theoretical support for the SEM-based causal inference methods that find the causal direction by comparing the marginal likelihood of the models in two directions; for an example of such methods, see [15].[8] One candidate model is given in Fig. 1(a), where X is exogenous for ψ (or $[\psi, (X, \psi)]$) operates a Bayesian cut in $p(X, Y, \theta, \psi)$, denoted by $\mathcal{M}_1$. The other corresponds to the factorization:

$$p(X, Y|\tilde{\theta}, \tilde{\psi}) = p(Y|\tilde{\theta})p(X|Y, \tilde{\psi}), \qquad (2)$$

where $[\tilde{\psi}, (Y, \tilde{\theta})]$ operates a Bayesian cut in $p(Y, X, \tilde{\theta}, \tilde{\psi})$, denoted by $\mathcal{M}_2$. Note that under the above models, the marginal likelihood of (X, Y) is the product of that of the conditioning variable and that of the conditional distribution. Ideally, if all the involved distributions are correctly specified, one would prefer the causal direction $X \to Y$ (resp. $Y \to X$) if $\mathcal{M}_1$ (resp. $\mathcal{M}_2$) gives a higher marginal likelihood.

Theorem 5. *Suppose that the two random variables X and Y are generated according to $\mathcal{M}_1$, and that the exogeneity-based causal model is identifiable. Let the prior distributions of the parameters be $p^*(\psi|\mathcal{M}_1)$ and $p^*(\theta|\mathcal{M}_1)$. For the given sample $(\mathbf{X}, \mathbf{Y})$, let $p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_1)$ be the marginal likelihood, i.e.,*

$$\begin{aligned}
&p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_1)\\
=&\prod_{i=1}^{N}\iint p(X_i, Y_i|\{\psi, \theta\})p^*(\theta|\mathcal{M}_1)p^*(\psi|\mathcal{M}_1)\mathrm{d}\theta\mathrm{d}\psi\\
=&\prod_{i=1}^{N}\int p(X_i|\theta)p^*(\theta|\mathcal{M}_1)\mathrm{d}\theta \cdot \prod_{i=1}^{N}\int p(Y_i|X_i, \psi)p^*(\psi|\mathcal{M}_1)\mathrm{d}\psi\\
=&\prod_{i=1}^{N} p(X_i|\mathcal{M}_1) \cdot \prod_{i=1}^{N} p(Y_i|X_i, \mathcal{M}_1).
\end{aligned}$$

Assume that by a one-to-one reparametrization we can represent $p(X, Y|\{\psi, \theta\})$ as $p(Y|\tilde{\theta})p(X|Y, \tilde{\psi})$, where Y is not exogenous for $\tilde{\psi}$. Let $p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_2)$ be the marginal likelihood of $\mathcal{M}_2$, i.e.,

$$\begin{aligned}
&p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_2)\\
=&\prod_{i=1}^{N}\iint p(X_i, Y_i|\{\tilde{\psi}, \tilde{\theta}\})p^0(\tilde{\theta}|\mathcal{M}_2)p^0(\tilde{\psi}|\mathcal{M}_2)\mathrm{d}\tilde{\theta}\mathrm{d}\tilde{\psi}\\
=&\prod_{i=1}^{N} p(Y_i|\mathcal{M}_2) \cdot \prod_{i=1}^{N} p(X_i|Y_i, \mathcal{M}_2),
\end{aligned}$$

where $\tilde{\theta}$ and $\tilde{\psi}$ have independent priors. As the sample size N goes to infinity, for any choice of $p^0(\tilde{\theta}|\mathcal{M}_2)$ and $p^0(\tilde{\psi}|\mathcal{M}_2)$, $p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_1)$ is always greater than $p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_2)$.

Proof. As the data were generated according to model $\mathcal{M}_1$, we have

$$\mathbb{E}\log p(X, Y|\mathcal{M}_1) = \int p(X, Y|\mathcal{M}_1)\log p(X, Y|\mathcal{M}_1)dxdy.$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}\log p(X, Y|\mathcal{M}_1) - \mathbb{E}\log p(X, Y|\mathcal{M}_2)\\
=&\int p(X, Y|\mathcal{M}_1)\log\frac{p(X, Y|\mathcal{M}_1)}{p(X, Y|\mathcal{M}_2)}dxdy\\
=&\mathcal{D}\big(p(X, Y|\mathcal{M}_1) \mid\mid p(X, Y|\mathcal{M}_2)\big),
\end{aligned}$$

where $\mathcal{D}(\cdot||\cdot)$ denotes the Kullback-Leibler divergence. Clearly the above quantity is non-negative, and it is zero if and only if $p(X, Y|\mathcal{M}_1) = p(X, Y|\mathcal{M}_2)$ for all possible x and y. However, this condition cannot hold, because the model $\mathcal{M}_1$ is assumed to be identifiable based on exogeneity.

Consequently, we have $\mathbb{E}\log p(X, Y|\mathcal{M}_1) > \mathbb{E}\log p(X, Y|\mathcal{M}_2)$. Moreover, according to the weak law of large numbers, as $N \to \infty$, $\frac{1}{N}\log p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_1)$ and $\frac{1}{N}\log p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_2)$ will convergence in probability to the quantities $\mathbb{E}\log p(X, Y|\mathcal{M}_1)$ and $\mathbb{E}\log p(X, Y|\mathcal{M}_2)$, respectively. That is, if N is large enough, $p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_1) > p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_2)$. □

However, the marginal likelihood depends heavily on the models or assumptions for the marginal and conditional distributions. Besides the exogeneity property, such approaches also make additional assumptions about the functions, such as structural constraints [21, 8, 27] and the smoothness assumption [15]. The proposed approach avoids such assumptions, by directly assessing the exogeneity property.

S2.1.1. A Simple Illustration on Parametric Models with Laplace Approximation

Here we use a somehow oversimplified parametric example to illustrate why the marginal likelihood implies the causal direction. Assume that $\mathcal{M}_1$ holds, that is, in factorization (1), X is exogenous to ψ. We will demonstrate that the likelihood for model (2) would be asymptotically smaller if we wrongly assume that Y is exogenous for $\tilde{\psi}$. We assume that there is a one-to-one correspondence between (θ, ψ) and $(\tilde{\theta}, \tilde{\psi})$. As seen from the proof of Theorem 5, the marginal distribution of (1) under $\mathcal{M}_1$ would be the same as that of (2) with the dependence between $\tilde{\theta}$ and $\tilde{\psi}$ taken into account. Suppose that the corresponding log marginal likelihood $\log p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_1)$, can be evaluated with the Laplace approximation in terms of $(\tilde{\theta}, \tilde{\psi})$ [12]:

$$\begin{aligned}
\log p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_1) \approx \log p(\mathbf{X}, \mathbf{Y}|\hat{\tilde{\theta}}, \hat{\tilde{\psi}}) &+ \log p^0(\hat{\tilde{\theta}}, \hat{\tilde{\psi}})\\
&- \frac{1}{2}\log|\Sigma_{\tilde{\theta},\tilde{\psi}}| + \frac{d}{2}\log(2\pi),
\end{aligned}$$

where $\hat{\tilde{\theta}}$ and $\hat{\tilde{\psi}}$ are the maximum a posterior (MAP) estimate, $p^0(\hat{\tilde{\theta}}, \hat{\tilde{\psi}})$ is the prior, $\Sigma_{\tilde{\theta},\tilde{\psi}}$ is the negative Hessian of $\log[p(\mathbf{X}, \mathbf{Y}|\tilde{\theta}, \tilde{\psi})p^0(\tilde{\theta})p^0(\tilde{\psi})]$ evaluated at $(\hat{\tilde{\theta}}, \tilde{\hat{\tilde{\psi}}})$, and d is the number of parameters.

On the other hand, under $\mathcal{M}_2$, the negative Hessian matrix becomes $\tilde{\Sigma}_{\tilde{\theta},\tilde{\psi}}$ which is block-diagonal and shares the same main diagonal block matrices $\Sigma_{\tilde{\theta}}$ and $\Sigma_{\tilde{\psi}}$ with $\Sigma_{\tilde{\theta},\tilde{\psi}}$. We then have $\log p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_1) - \log p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_2) \approx \frac{1}{2}\big(\log|\tilde{\Sigma}_{\tilde{\theta},\tilde{\psi}}| - \log|\Sigma_{\tilde{\theta},\tilde{\psi}}|\big) = \frac{1}{2}\big(\log|\Sigma_{\tilde{\theta}}| + \log|\Sigma_{\tilde{\psi}}| - \log|\Sigma_{\tilde{\theta},\tilde{\psi}}|\big)$. One can show that $|\Sigma_{\tilde{\theta},\tilde{\psi}}| < |\Sigma_{\tilde{\theta}}| \cdot |\Sigma_{\tilde{\psi}}|$ if $\Sigma_{\tilde{\theta},\tilde{\psi}}$ is not block-diagonal; for a proof, see [1, page 239]. Hence, we have $\log p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_1) > \log p(\mathbf{X}, \mathbf{Y}|\mathcal{M}_2)$ asymptotically.

S2.2. Relation to Invariance of SEMs

The proposed bootstrap-based method provides a way to examine if an equation is structural or not. Suppose $Y =$

generally speaking, the above error-based definition is consistent with Definition 1 only when the functional class is well constrained. Otherwise, if the function and the distribution of the assumed cause are related in some way, the above definition is not rigorous.

[8] Note that due to computational difficulties, this method doe snot evaluate the marginal likelihood, but approximate it wiht the maximum regularized likelihood.

$f(X, E)$, where $E \perp\!\!\!\perp X$, is a structural causal model in that f is invariant to changes in the distribution of X [18]. One can then see that since E and X are independent processes, the bootstrapped $\hat{P}^{*(b)}(X)$ is independent from the underlying $\hat{p}^{*(b)}(E)$, and hence independent from $\hat{p}^{*(b)}(Y|X) = \hat{p}^{*(b)}(E)/\left|\frac{\partial f}{\partial E}\right|$.

Now consider the other direction. According to [9], we can always find an equation $X = \tilde{f}(Y; \tilde{E})$ such that $\tilde{E} \perp\!\!\!\perp Y$; suppose this equation is not structural, in that $\tilde{f}$, or in particular, $\left|\frac{\partial \tilde{f}}{\partial \tilde{E}}\right|$ is dependent on $p(Y)$. Again, we have $\hat{p}^{*(b)}(X|Y) = \hat{p}^{*(b)}(\tilde{E})/\left|\frac{\partial \tilde{f}}{\partial \tilde{E}}\right|$. The bootstrapped $\hat{p}^{*(b)}(Y)$ and $\hat{p}^{*(b)}(X|Y)$ are then dependent due to the dependence between $\left|\frac{\partial \tilde{f}}{\partial \tilde{E}}\right|$ and $\hat{p}^{*(b)}(Y)$.

In particular, the SEM-based causal inference approaches [21, 8, 27, 15] constrain the functions f to be simple in respective senses; consequently they are not so flexible as to change with the input distribution $p(X)$, and then the independence between the input X and the noise E serves as a surrogate to achieve the exogeneity condition of X for the parameters in $p(Y|X)$.

Compared to SEM-based approaches, the proposed exogeneity-based approach avoids the constraints on the functional causal model f. On the other hand, some SEM-based approaches have clear identifiability conditions under which the reverse direction $Y \to X$ that induces the same joint distribution on (X, Y) does not exist in general, given the causal direction $X \to Y$; for instance, see [8, 27]. However, to find theoretical identifiability results for the proposed approach, one has to establish the identifiability conditions in terms of data distributions, which turns out to be extremely difficult.

Index of Authors