Deontic Logic and Normative Systems

16th International Conference, DEON 2023

Editors
Juliano Maranhão
Clayton Peterson
Christian Straßer
Leendert van der Torre

Contents

Editorial Preface

The biennial DEON conferences are designed to promote interdisciplinary cooperation amongst scholars interested in linking the formal-logical study of normative concepts and normative systems with computer science, artificial intelligence, philosophy, ethics, linguistics, organization theory and law.

There have been fifteen previous DEON conferences held in Amsterdam (December 1991), Oslo (January 1994), Sesimbra (January 1996), Bologna (January 1998), Toulouse (January 2000), London (May 2002), Madeira (May 2004), Utrecht (July 2006), Luxembourg (July 2008), Fiesole (July 2010), Bergen (July 2012), Ghent (July 2014), Bayreuth (July 2016), Utrecht (July 2018), and Munich (July 2021). The 16th occurrence of the conference, DEON2023, was held for the first time in North America at Université du Québec à Trois-Rivières.

DEON2023 focused on the theme *Theoretical and technical limitations of automated behavior*. This theme aimed to address the growing literature in machine ethics attempting at creating ethical machines through AI and machine learning. This, in conjunction with the fact that there is a tendency to anthropomorphise AI, has led some scholars to believe that we should thrive to define and build machines that would be better than humans at making ethical choices and behaving ethically. While many limitations to such attempts are known, including the framing problem, incompatible ethical theories, conflicts between rules, and computation time, it remains that many problems and characteristics well known in the deontic logic literature (e.g. deontic paradoxes, conflicting norms and obligations) and that have a direct impact on the mere possibility of defining truly autonomous ethical machines seem to have been overlooked in the machine ethics literature. As such, DEON2023 aimed to raise awareness on important aspects from the deontic logic literature that should impact machine ethics by focusing on the theoretical and technical limitations of automated behavior.

DEON2023 was organized in the context of a partnership and collaboration between Clayton Peterson (Université du Québec à Trois-Rivières) and Leendert van der Torre (Université du Luxembourg) through their INTEGRAUTO Audace International funded project. The conference was organized as a hybrid conference, with live transmission and recordings of the conference, but with speakers required to be physically present (i.e. in person attendance was required for speakers). Talks, posters, and poster teasers were made publicly available on the conference's website.

The conference was divided into a main track and a poster track. Papers were first accepted or rejected for the proceedings, and then, for the papers accepted for the proceedings, papers were either accepted for oral presentation or for poster presentation. This decision was based on many criteria, including whether papers are of interest to a larger or smaller community, while trying to be inclusive and give the authors the opportunity to present and discuss their work. We received 43 submissions, out of which 27 materialized as full papers, and 6 submissions to the poster track. Out of these 27 papers, 10 were included as contributed talks, whereas 7 where redirected to the poster track (with full papers published in the proceedings). Each paper was carefully evaluated by at least three reviewers during a blind review process. Overall, DEON2023 presented 5 keynote speakers, 10 contributed talks, and 12 posters, with 17 papers and 5 abstracts published in the conference proceedings.

DEON2023

UQTR
Université du Québec
à Trois-Rivières

CNIMI Centre national intégré
du manufacturier intelligent
Propulsé par l'UQTR et le Cégep de Drummondville

Laboratoire d'Intelligence
Artificielle Appliquée - LI2A
UQTR
Université du Québec
à Trois-Rivières

Fièrement
propulsé
par les FRQ
Québec

Fonds National de la
Recherche Luxembourg

Conseil de recherches en
sciences humaines du Canada

Social Sciences and Humanities
Research Council of Canada

Canada

Keynote speakers

Marina De Vos (University of Bath)

Marina De Vos is a senior lecturer/associate professor in artificial intelligence and the director of training for the UKRI Centre for Doctoral Training in Accountable, Responsible, and Transparent AI at the University of Bath. With a strong background in automated human reasoning, Marina's research focuses on enabling improved access to specialist knowledge, the logical foundations of AI systems, explainable artificial intelligence methods, and modelling the behaviour of autonomous systems. In her work on normative multi-agent systems, Marina combines her interests in the development of software tools and methods, drawing from a diverse range of domains including software verification, logic programming, legal reasoning, and AI explainability, to effectively model, verify and explain autonomous agents. Currently, Marina's exploration involves systems that possess the ability to autonomously evolve through external and internal stimuli.

Title From normative systems to business rules

Abstract In human society, norms, policies, and laws serve as mechanisms to describe, guide, and regulate expected behaviour. These rules outline the desired conduct of individuals and specify rewards or penalties for compliance or violation. Similarly, these concepts can be applied to socio-technical systems, encompassing both human and software agents. Within such normative systems, agents possess autonomy, enabling them to decide whether to adhere to norms or deviate from them. This talk delves into the representation and computational reasoning of norms, policies, and laws, while ensuring sufficient clarity for human participants involved in the system. The focal point is InstAL, a domain-specific language (DSL) designed to capture deontic concepts and the effects of agents' actions. Execution is facilitated through answer set programming, a declarative logic programming language. Together, they provide a powerful approach for modelling, verifying, monitoring, and revising norms within socio-technical systems. To illustrate the practical implications, this presentation presents a case study that explores the compliance of business processes with specific aspects of the General Data Protection Regulation (GDPR). To seamlessly integrate with established practices in business process modelling and semantic web technology, an ODRL layer was developed on top of InstAL, facilitating smooth integration with existing workflows. This talk aims to offer valuable perspectives on the potential of combining InstAL, answer set programming, and the ODRL layer to effectively model, verify, monitor, and revise norms within socio-technical systems, as exemplified through the GDPR compliance case study.

Huimin Dong (Sun Yat-Sen University)

Huimin Dong, an Assistant Professor at Sun Yan-sen University's Department of Philosophy (Zhuhai), specializes in developing formal models for normative reasoning. Her interdisciplinary research covers logic, philosophy, ethics, law, and AI, with a particular focus on deontic logic, nonmonotonic reasoning, and logic-based methods for AI ethics and law.

Title Resolving the Paradox of Free Choice Permission: A Semantic Approach

Abstract The concept of free choice permission is commonly understood as follows: if it is permitted to do α or β, then it is permitted to do α and it is permitted to do β. This differs from a permission that simply implies the absence of prohibition. However, when applying monotonic reasoning to this type of permission, a permission to do α logically leads to a permission to do both α and β. Problems arise when we introduce a prohibition on doing β, resulting in a paradox of free choice permission. Various proof theory solutions have been proposed to address the nonmonotonic aspects of this issue. In this talk, I will present a semantic approach aimed at resolving this problem. Following the tradition of dynamic logic, I will focus on the concept of open reading as the semantic core of free choice permission. Furthermore, I will discuss how the inclusion of normality can be incorporated into this framework to resolve the free choice paradox.

Lou Goble (Willamette University)

Lou Goble studied philosophy and logic at Oberlin College (B.A.) and the University of Pittsburgh (M.A., Ph.D.), where he worked with Wilfrid Sellars, Alan Ross Anderson, and Nuel D. Belnap, Jr., amongst others. He then taught philosophy at the University of Wisconsin, Madison, and the University of North Carolina, Chapel Hill, before retiring for a while to the wilds of Oregon. He emerged from the shadows to teach at Willamette University, until his full retirement. The author of The Kalevide (a novel) and editor of The Blackwell Guide to Philosophical Logic and, with J. J. Ch. Meyer, Deontic Logic and Artificial Normative Systems (Proceedings of

DEON 2006), he continues to write, though not usually for publication, on questions in philosophical logic and the philosophy of language, and more or less academic matters.

Title Preemption and Plausible Oughts

Abstract I present a problem for deontic logic, a puzzle to invite further investigation. It is the problem of preemption: Imagine a case in which (a) there is something, X, that might occur but really should not, but also (b) there is something else, Y, that, if done, would preempt X, i.e., if Y were done, then X would not occur, while (c) if Y were not done, then X would happen. Moreover, in this case, neither X nor Y is determined; it is even possible (d) that Y not be done and X not happen, though, in light of (c), that is a remote possibility, even far-fetched. We may suppose too (e) it would be better for Y to be done and X not occur than for Y not to be done and X occur. And yet, although it is a remote possibility, (f) having Y not be done and X not occur would be better still than for Y to be done and X not occur. Given (a)–(f), and especially (a), (b), (c) and (e), the inference to (g) that Y should be done, seems clear, despite (d) and (f). The problem for deontic logic is to explain the validity of that inference in a plausible way. I present four approaches to that problem, all within the framework of branching time structures designed to accommodate the indeterminism inherent in the case. One of these approaches is familiar but it fails to account for the inference. The others are offered as more realistic alternatives that work better. Yet each has its drawbacks, and there is more work to be done. My primary purpose is to encourage that research.

John Horty (University of Maryland)

John Horty received his BA in Classics and Philosophy from Oberlin College and his PhD in Philosophy from the University of Pittsburgh; he is currently a Professor of Philosophy at the University of Maryland with affiliate appointments in Computer Science and the Institute for Advanced Computer Studies. His interests include logic, artificial intelligence, ethics, epistemology, philosophy of language, and philosophy of law. John Horty is the author of four books as well as papers on a variety of topics in logic, philosophy, and computer science. His work has been supported by three fellowships from the National Endowment for Humanities and several grants from the National Science Foundation, by visiting fellowships at the Netherlands Institute for Advanced Studies and at the Center for Advanced Studies in Behavioral

Sciences at Stanford University, and more recently, by a Humboldt Research Award.

Title Knowledge representation for computational normative reasoning

Abstract I will talk about issues involved in designing a machine capable of acquiring, representing, and reasoning with information needed to guide everyday normative reasoning - the kind of reasoning that robotic assistants would have to engage in just to help us with simple tasks. After reviewing some current top-down, bottom-up, and hybrid approaches, I will define a new hybrid approach that generalizes ideas developed in the fields of AI and law and legal theory. Joint work with Ilaria Canavotto.

Réka Markovich (Université du Luxembourg)

Réka Markovich researches computational legal theory and studies its applications in Artificial Intelligence and legal reasoning. Her focus areas are legal knowledge representation, normative multi-agent systems, deontic logic, machine ethics, and XAI. Réka has an interdisciplinary background: she has degrees in law, in logic, and in communications, and a PhD in logic. Réka is currently an independent research scientist at the Department of Computer Science at the University of Luxembourg where she is the head of the newly established Computational Law and Machine Ethics (CLAiM) group in the Interdisciplinary Lab for Intelligent and Adaptive Systems. She represents Luxembourg on the board of the Benelux Association for AI and in 2021, she got elected to the international Steering Committee of the Foundation for Legal Knowledge Based Systems.

Title A formal theory of rights

Abstract Deontic logics traditionally concern impersonal obligations, however, to understand some basic concepts and structures of law, agents must be explicitly taken into account. For understanding rights, one actually needs to consider pairs of agents and formally map the variants of relations between them. In the talk, I give a comprehensive overview of the formal theory of rights I have been working on in the last few years. This work contributes to the tradition of the theory of normative positions, but exceeds that by defining formal characterizations using a multi-modal language, by sketching a theory of legal metaphysics, and also by offering resolution to difficulties rights theories often face.

Chairs of the program committee

Juliano Maranhão (Universidade São Paulo)
Christian Straßer (Ruhr-Universität Bochum)
Leendert van der Torre (Université du Luxembourg)

Chair of the local organizing committee

Clayton Peterson (Université du Québec à Trois-Rivières)

Local organizing committee

Ismaïl Biskri (Université du Québec à Trois-Rivières)
Marc-André Gaudreau (Université du Québec à Trois-Rivières)
Sousso Kelouwani (Université du Québec à Trois-Rivières)
Clayton Peterson (Université du Québec à Trois-Rivières)

Program committee

Christoph Benzmüller (Otto-Friedrich-Universität Bamberg)
Ismaïl Biskri (Université du Québec à Trois-Rivières)
Jan Broersen (Utrecht University)
Mark A. Brown (Syracuse University)
Ilaria Canavotto (University of Maryland, College Park)
Fabrizio Cariani (University of Maryland, College Park)
Jose Carmo (University of Madeira)
Huimin Dong (Sun Yat-Sen University)
Federico L.G. Faroldi (University of Pavia)
Guido Governatori (Independent researcher)
Aleks Knoks (University of Luxembourg)
Piotr Kulicki (John Paul II Catholic University of Lublin)
Réka Markovich (Université du Luxembourg)
Alessandra Marra (MCMP - LMU Munich)
Paul McNamara (UNH)
Robert Mullins (University of Oxford)
Xavier Parent (TU Wien)
Gillman Payette (University of Calgary)
Gabriella Pigozzi (Université Paris-Dauphine)
Martin Rechenauer (Ludwig Maximilian University of Munich)

Antonino Rotolo (University of Bologna)
Olivier Roy (Universität Bayreuth)
Chenwei Shi (Tsinghua University)
Allard Tamminga (University of Greifswald)
Paolo Turrini (University of Warwick)
Kees van Berkel (Ruhr-University Bochum)
Frederik Van De Putte (Erasmus University of Rotterdam & Ghent University)
Peter Vranas (University of Wisconsin-Madison)
Malte Willer (University of Chicago)
Tomoyuki Yamada (Hokkaido University)

DEON Steering committee

Jan Broersen (Utrecht University) - Chair
John Horty (University of Maryland) - Vice Chair
Christoph Benzmüller (Freie Universität Berlin)
Cleo Condoravdi (Stanford University)
Melissa Fusco (Colombia University)
Beishui Liao (Zhejiang University)
Juliano Maranhão (University of São Paulo)
Alessandra Marra (Ludwig-Maximilians-Universität München)
Paul McNamara (University of New Hampshire)
Joke Meheus (Ghent University)
Gabriella Pigozzi (Université Paris-Dauphine)
Paul Portner (Georgetown University)
Antonino Rotolo (University of Bologna)
Oliver Roy (Universität Bayreuth)
Leendert van der Torre (University of Luxembourg)
Malte Willer (University of Chicago)

Acknowledgments

We gladly acknowledge financial support from *Fonds de recherche du Québec*, *Fonds National de Recherche du Luxembourg*, Social Sciences and Humanities Research Council, as well as from Université du Québec à Trois-Rivières. We would further like to thank College Publications for their support in publishing DEON2023's proceedings.

DEON2023 was funded by the *Fonds de recherche du Québec* [FRQSC-AUDACE-310754] and the *Fonds National de Recherche du Luxembourg* [INTEGRAUTO INTER/AUDACE/21/16695098].

DEON2023 was further supported in part by funding from the Social Sciences and Humanities Research Council [Institutional grant - Exchange].

DEON2023 also received financial support from the Université du Québec à Trois-Rivières' *Fonds d'animation de la recherche* [FAR - UQTR].

Rights and Practical Reasoning in Deontic Logic

Huimin Dong [1]

Institute of Logic and Cognition, Sun Yat-sen University, China

Dragan Doder

Department of Computer Science, Utrecht University, the Netherlands

Xu Li, Réka Markovich, Leendert van der Torre

Department of Computer Science, University of Luxembourg, Luxembourg

Marc van Zee

Google Research, Denmark

Abstract

This paper brings together two traditions in deontic logic: the theory of normative positions, that is, reasoning about different types of rights, and practical reasoning, which has special relevance from the viewpoint of artificial intelligence (AI). We do this by exploring the role epistemic rights play in practical reasoning. Rights such as the right to know are intended to enable us to make informed decisions. They often play a role in determining what kind of plans we can make. A patient has the right to know his hospital test results so he can choose his treatment after his doctor has fulfilled her duty of informing him about possible risks and outcomes. This paper investigates, from the "database perspective", the role of (epistemic) rights in planning different scenarios from the database perspective and the dynamics of temporal beliefs and intentions. We take this perspective, extend the logic with deontic notions, and illustrate this with a running example.

Keywords: normative reasoning in AI, practical reasoning, normative positions

1 Introduction

Research in deontic logic includes a decades-long investigation into normative positions, benchmarked by Sergot's chapter in the first volume of the Handbook of Deontic Logic [24]. From the perspective of artificial intelligence

[1] Corresponding author: `donghm.logic@gmail.com`

(AI), practical reasoning is one of the most important topics in deontic logic and normative reasoning, benchmarked by Thomason's chapter in the second volume of the Handbook of Deontic Logic [27]. However, the topics of normative positions and practical reasoning are hardly ever brought together. The aim of this paper is to bring practical reasoning as used in AI to the field of deontic logic, with a special focus on the use of reasoning with rights from the database perspective [26,28].

In the tradition of reasoning about rights, the logics developed for normative positions (by Kanger [13] and Lindahl [15]) were initially aimed at mapping the space of logically possible legal relations between two given agents, differentiating between more and more variants [24]. These logics used a very weak action logic (Chellas called this system ET [7]) preventing the derivation of extensive consequences. Several more recent papers focusing on the conceptual elaboration of different notions of right like [17] adopt this approach. These logics thus have limited use for representing how an agent can reason practically about its actions in detail based on its own normative positions.

In contrast, most research on practical reasoning disregards rights and normative relations. BDI (Belief-Desire-Intention) logics (e.g., [8,22,28]) focus on specifying the relations between various mental states such as belief, desire, intention, and goal, but they traditionally ignore normative concepts. BOID (Belief-Obligation-Desire-Intention) [5] later incorporated obligations, but did not do so with normative positions.

This paper contributes to closing this gap by pointing out that in everyday life, we plan our actions by deliberating different scenarios. Our rights can play an important role in this planning, for instance when we come up with an optimal scenario where we have the right to do or get what we want. We start from the "database perspective" [26], a recent proposal that differentiates between a planner and belief-intention databases. The planner is engaged in some form of (temporal) practical reasoning, and in this process updates the databases. The task for the databases is to remain coherent. Van Zee *et al.* [28] formalized the databases using (Par)ameterized-time Action Logic (PAL) logic and providing AGM-like (Alchourrón-Gärdenfors-Makinson) postulates [1] for the revision of beliefs and intentions. Our main research question is: "how to characterize (epistemic) rights in terms of the role they play in practical reasoning". This is broken down into the following three sub-questions:

RQ1: the role and components of rights in practical reasoning;

RQ2: how to extend PAL [28] with the concepts needed for (epistemic) rights in practical reasoning;

RQ3: how to use this formal framework to model (epistemic) rights in practical reasoning.

We will characterize some variants of the right to know—with an emphasis on power—in terms of how they influence the dynamics of planning from the database perspective". The approach we use contributes to several aspects compared to previous research on the dynamic nature of normative positions (i.e., frameworks on the power type of right). For instance, from a conceptual point of view, it emphasizes the practical reasoning aspect, while from a technical point of view, it expresses the dynamics by using two revision operators as two kinds of coherence on a database. We will discuss this in detail at the end of the paper.

The layout of this paper follows the three research questions. In Section 2, we discuss the role of epistemic rights in practical reasoning and introduce the logic of intentions [28]. We extend the logic of intentions with obligation and permission in Section 3, and in Section 4 we apply the new logic to develop a revision operator to characterize Hohfeldian power. Section 5 ends the paper with our conclusions and future work.

2 Background

This section provides the background for this paper. First, we provide a short introduction to the theory of rights within the theory of normative positions, then we introduce a running example that we formalize throughout the paper. Finally, we give a summary of the database perspective.

2.1 (Epistemic) Rights in Deontic Logic: Theory of Normative Positions

The theory of normative positions in deontic logic refers to the tradition of formalizing normative relations between pairs of agents and their resultant relative positions. The theory relies on different meanings of the word "right" and their correlative duties put forward by Hohfeld. The tradition began with the work of Kanger and Kanger [13] and Lindahl [15], and has been developed by many others (e.g., [16,12,17,9]) more recently. The basic idea is that "right" can have different meanings, and the four atomic ones—in Hohfeldian terminology—are claim-right, privilege, power, and immunity. Each comes with its own correlative duty. That is, whenever an agent has one of these right positions, the counterparty has a duty position: duty (in the narrow sense), no-claim, liability, or disability, respectively. Claim-right is a claim that the duty bearer should take a particular action. Duty is the directed version of the classical notion of obligation in deontic logic. Privilege refers to the freedom of the right-holder to take a particular action when the counterparty has no claim to refrain him from doing so. This is the relationalized version of a weak permission. Power is when the right-holder has the possibility of changing the counterparty's normative positions with a special action. If a professor has the right to hand out homework, that means that she can create a duty for her students to do their homework. Immunity means that the counterparty does not have the power to change the right-holder's normative position. The interpretation of epistemic rights with Hohfeldian categories was put forward by the epistemologist Lani Watson [29], and the logical formalization of this interpretation has been articulated in some recent papers [18,19,14]. The formalization of the right to know in [18] and [19] uses the weak action logic referred to above.

2.2 Running Example

Avery (also called “Patient”, or simply P) suspects he has an illness that makes him eligible for early retirement, but he doesn’t want to apply until he is sure. He intends to get tested, knowing that he has the right to know the results as this is one of the patient’s listed rights under the law. After the tests, Avery exercises his right to know by asking for the results. The ‘right to know’ is understood as a Hohfeldian power by default. When Avery asks for the results, this puts an obligation on the doctor to inform him. That obligation means that Avery’s right to know becomes a claim-right. The doctor may intend to ignore the request, violating her obligation. This could make Avery give up his plan to apply for early retirement. Or he could decide to complain to the hospital director with the expectation that he would then get the results. After all, he still believes he has the right to be informed.

2.3 The Database Perspective

The logic of intentions has been studied in the fields of theory of mind and artificial intelligence. Van Zee et al. [28] provided a logic for reasoning about the dynamics of intentions and beliefs in time, formalizing Shoham’s database perspective [26]. This approach uses a temporal branching time logic called Parameterized-time Action Logic (PAL).

Definition 2.1 [The PAL Language] Let $\mathsf{Act} = \{a, b, c, \dots\}$ be a finite set of deterministic primitive actions and let $\mathsf{Prop} = \{p, q, r, \dots\} \cup \{pre(\bar{a}), post(a)\}$ be a finite set of propositions where $\bar{a} = (a_1, a_2, \dots)$ is a non-empty sequence of actions and $\{a, a_1, a_2, \dots\} \subseteq \mathsf{Act}$ are actions. The language $\mathscr{L}$ of the logic is as follows:

$$\varphi ::= \chi_t \mid do(a)_t \mid \Box_t\varphi \mid \varphi \wedge \varphi \mid \neg\varphi,$$

where $\chi \in \mathsf{Prop}$, $a \in \mathsf{Act}$, and $t \in \mathbb{N}$.

Intuitively, p_t means that p is true at time t, and $do(a)_t$ means that action a is executed at time t. Then, $pre(\bar{a})_t$ means that the precondition for a sequence of actions $\bar{a} = (a_1, ..., a_n)$ at time t is satisfied. Preconditions are defined on action sequences to ensure that it is possible to do the all the intended actions together (see the original paper [28] for more details). For instance, $pre(a_1, a_2, a_3)_0$ indicates that the precondition for doing a_1 at time 0, a_2 at time 1 and a_3 at time 2 is true. Then, $post(a)_t$ represents the postcondition for a at time t. The modal operator $\Box_t$ is interpreted as a *temporal* necessity for the planner, so a formula of the form $\Box_t p_{t'}$ means “it is necessary at time t that p is true at time t'”. This necessity means that no matter which actions are executed between time t and time t', p will hold in t'.

This provides a sound and strongly complete axiomatization. Due to space constraints, we only provide axioms relevant to this paper.

Definition 2.2 [Axiomatization (Partial)] Here are some PAL axioms. The full axiomatization can be found in the work of Van Zee [28], Section 2.3.

(A5) $\Box_t\varphi \to \Box_{t+1}\varphi$ (A8) $do(a)_t \to post(a)_{t+1}$

(A6) $\bigvee_{a\in\mathsf{Act}} do(a)_t$ (A9) $pre(a)_t \to \Diamond_t do(a)_t$

(A7) $do(a)_t \to \neg do(b)_t$, where $b \neq a$ (NEC) From φ, infer $\Box_0\varphi$

The intuitive meaning of some of the above axioms is explained below. Axiom A5 indicates continuity along the progression of time. If something is necessary at time t, then it remains necessary at the next time point $t+1$. Axiom A6 and Axiom A7 together state that at any given time point, one and only one action can be executed.

Due to space constraints, we omit the technical details of the semantics, and provide a short description only (see the work of Van Zee *et al.* [28], Section 2.2, for full details). PAL semantics is similar to that of computation tree logic (CTL)* [23] except that each transition between two consecutive states, the transition is also labeled by an action. A model (T, π) consists of a tree T and a path π. Trees have their root at time 0. Then, $T, \pi \models p_t$ means that proposition p is in the valuation function of the state corresponding to path π at time t (denoted as π_t). It follows that $T, \pi \models do(a)_t$ means that the transition from state π_t to π_{t+1} is labeled with action a. And $T, \pi \models \Box_t\varphi$ means that φ is true for all paths that are equivalent to π up to time t (i.e., they have the same states as π up to time t) in tree T. The other truth definitions are defined as per usual.

Note that the semantics distinguishes regular or strong beliefs from weak beliefs, which are beliefs contingent on the intended actions. The set $\mathbb{SB}$ of all *strong beliefs* is generated by Boolean combinations of $\Box_0\varphi$ where $\varphi \in \mathscr{L}$. A strong belief is an element of $\mathbb{SB}$. A set SB of strong beliefs is the deductive closure of a subset of $\mathbb{SB}$ such that $SB = Cn(\Sigma)$ where $\Sigma \subseteq \mathbb{SB}$. Semantically, a strong belief is a formula that is true for all the paths of the tree, meaning that they are independent of a specific future or *plan* (i.e., a specific sequence of intentions).

Definition 2.3 [Belief-Intention Database] A *belief-intention database* (SB, I) consists of a belief database SB and an intention database I:

- $SB \in \mathbb{SB}$ is a set of strong beliefs closed under consequence: $SB = Cn(SB)$;
- $I = \{(a_1, t_1), (a_2, t_2), \dots\}$ is a set of intentions (a_i, t_i) where $a_i \in \mathsf{Act}$ such that no two intentions exist at the same time point, i.e., if $i \neq j$ then $t_i \neq t_j$.

Weak beliefs are obtained by adding intentions to the strong beliefs and closing the result under consequence. Thus, a weak belief is closely related to a contingent or specific *plan*.

Definition 2.4 [Weak Beliefs] Given a belief-intention database (SB, I), weak beliefs are defined as follows:

$$WB(SB, I) = Cn(SB \cup \{do(a)_t | (a, t) \in I\}).$$

Commitment to intentions is characterized using a coherence condition stating that it is possible to perform all the intended actions.

Definition 2.5 [Coherence] Given an intention database $I = \{(b_{t_1}, t_1), \dots, (b_{t_n}, t_n)\}$ with $t_1 < \dots < t_n$, let

$$Cohere(I) = \Diamond_0 \bigvee_{\substack{a_t \in Act : t \notin \{t_1,\ldots,t_n\} \\ a_t = b_t : t \in \{t_1,\ldots,t_n\}}} pre(a_{t_1}, a_{t_1+1}, \ldots, a_{t_n})_{t_1}.$$

More precisely, when we have a set of intended actions at non-consecutive time points $t_1, \ldots, t_n$, it is always possible at the initial time point 0 to carry out these intended actions by incorporating additional actions in the remaining time points. We say that a given belief-intention database (SB, I) is *coherent* iff SB is consistent with $Cohere(I)$, i.e., $SB \nvdash \neg Cohere(I)$.

A proposition relating weak beliefs to coherence [28] is shown below.

Proposition 2.6 *Given a belief-intention database* (SB, I)*, if* (SB, I) *is coherent, then* $WB(SB, I)$ *is consistent.*

Revision operators are then defined for both beliefs and intentions. The ones presented here are almost the same as those of Van Zee *et al.* but are slightly simpler. [2]

Definition 2.7 [Intention revision function] An *intention revision function* $\otimes$ maps a belief-intention database and an intention to a belief-intention database such that

$$(SB, I) \otimes i = (SB, I'),$$

where the following postulates hold:
$(P1)$ (SB, I') is coherent;
$(P2)$ If $(SB, \{i\})$ is coherent, then $i \in I'$;
$(P3)$ If $(SB, I \cup \{i\})$ is coherent, then $I \cup \{i\} \subseteq I'$;
$(P4)$ $I' \subseteq I \cup \{i\}$;
$(P5)$ For all I'' with $I' \subset I'' \subseteq I \cup \{i\}$:$(SB, I'')$ is not coherent.

Postulate (P2) states that new intention i takes precedence over all other current intentions. If possible, it should be added even if all current intentions must be discarded. Postulate (P3) and (P4) together state that if it is possible to simply add the intention, then this is the only change that is made. These two postulates are comparable to the inclusion and vacuity of AGM. Finally, (P5) states that we do not discard intentions unnecessarily.

Definition 2.8 [Belief revision function] A *belief revision function* $\circ$ maps a belief-intention database and a strong belief formula φ to a belief-intention database such that

$$(SB, I) \circ \varphi = (SB', I'),$$

where:

- SB' is the result of revising SB with a φ that satisfies the AGM postulates [1],

[2] Our revision operators differ from those of Van Zee *et al.* in three ways. 1) They bind their revision operators up to a time point t, which is a mere technical detail to prove a representation theorem, so we leave this out. 2) For technical reasons, they represent a belief set SB as a propositional formula ψ such that $SB = \{\varphi | \psi \vdash \varphi\}$, but we simply use SB directly. 3) They define a revision operator for revising with the pair (φ, i), which is slightly more general than our variant but is used only for edge cases.

- I' is the result of revising the new beliefs with the empty intention ϵ so that coherence is restored, i.e., $(SB', I) \otimes \epsilon = (SB', I')$.

Note that, by this definition, the revision of strong beliefs cannot be triggered by intention revision, but it can trigger intention revision. Intuitively, this makes sense: one would not wish to change one's strong beliefs after adopting an intention, but might want to update one's intentions after learning new information.

3 Formalizing Obligation (and Claim-Right)

In this section, we formalize obligation—and thus also claim-right, its corresponding notion in the normative position theory—from the database perspective, while not extending them in any way. It turns out that we are able to model these concepts quite naturally using only beliefs and intentions. In the next section, we extend the coherence condition so that we are able to use deontic notions when revising with new information.

We model the doctor-patient example with only one belief base and one intention base. In this case, the beliefs may be seen as shared or common beliefs, and the intentions may be seen as shared intentions. [3]

In our minimal formalization, we introduce only some special actions such as test, ask, and inform. And we introduce only some special propositions such as pre/postconditions for actions and a violation constant for obligations.

3.1 Actions

To model the action that agent i informs agent j about proposition p, we use the action $inform(i, j, p)$. And we use $\textit{is-informed-whether}(i, p)$ propositions (abbreviated as $iiw(i, p)$) to model whether agent i is informed about the truth or falsehood of p. We have that $iiw(i, p)$ is a precondition of $inform(i, j, p)$ and $iiw(j, p)$ is a postcondition of that action.

We assume that the doctor can not only learn whether p is true or false by being informed but can also carry out tests to find out. So $test(i, p)$ has postcondition $iiw(i, p)$.

Since p (whether the patient is ill) is the focus and is always repeated in our running example, we simplify things by omitting it from the actions and propositions below.

Example 3.1 [Running example] Let $\mathsf{Prop} = \{iiw(\mathtt{D}), iiw(\mathtt{P}), \mathtt{v}\}$ and $\mathsf{Act} = \{test(\mathtt{D}), ask(\mathtt{P}, \mathtt{D}), inform(\mathtt{D}, \mathtt{P}), ignore(\mathtt{D}, \mathtt{P}), complain(\mathtt{P}, \mathtt{HD})\}$. These are interpreted as follows:

- $test(\mathtt{D})$: the doctor tests whether the patient is ill;
- $ask(\mathtt{P}, \mathtt{D})$: the patient asks the doctor whether he is ill or not;
- $inform(\mathtt{D}, \mathtt{P})$: the doctor informs the patient whether he is ill or not;

[3] Note that this means that the revision operators aren't revision operators for a particular agent but for the entire system. Thus, if we revise intentions related to a particular agent, this may affect the intentions of other agents.

- $ignore(\mathtt{D},\mathtt{P})$: the doctor ignores the patient's request;
- $complain(\mathtt{P},\mathtt{HD})$: the patient complains to the director;
- $iiw(\mathtt{D})$ / $iiw(\mathtt{P})$: the doctor/patient is informed whether the patient is ill;
- v: a violation occurs.

While PAL defines pre- and postconditions as primitive propositions, we introduce the following abbreviations in our running example:

(i) $post(test(\mathtt{D})) = iiw(\mathtt{D})$: after the doctor has carried out the tests, she knows whether the patient is ill or not;

(ii) $pre(inform(\mathtt{D},\mathtt{P})) = iiw(\mathtt{D})$: the doctor can only inform the patient if she knows whether the patient is ill or not;

(iii) $post(inform(\mathtt{D},\mathtt{P})) = iiw(\mathtt{P})$: after the doctor has informed the patient, he knows whether he is ill or not;

(iv) $post(ask(\mathtt{P},\mathtt{D})) = pre(ignore(\mathtt{D},\mathtt{P}))$: the doctor can only ignore the request if the patient has made the request.

(v) $post(ask(\mathtt{P},\mathtt{D})) = pre(inform(\mathtt{D},\mathtt{P}))$: the doctor can inform the patient whether he is ill or not upon request;

(vi) $post(ignore(\mathtt{D},\mathtt{P})) = pre(complain(\mathtt{P},\mathtt{HD}))$: the patient can only complain to the director if the doctor ignores his request.

While PAL defines preconditions for action sequences as primitive propositions, we use the following inductive definition so that we can also include the precondition formulas above in preconditions for action sequences: $pre(a,\bar{b})_t = pre(a)_t \wedge \Diamond_t(do(a)_t \to pre(\bar{b})_{t+1})$.

We can use PAL axiomatization (Def. 2.2) and the above formulas to derive new formulas:

- $do(test(\mathtt{D}))_t \to iiw(\mathtt{D})_{t+1}$ (A8, (i));
- $\Box_0(do(inform(\mathtt{D},\mathtt{P}))_t \to iiw(\mathtt{P})_{t+1})$ (A8, (iii), NEC);
- $iiw(\mathtt{D})_t \to \Diamond_t do(inform(\mathtt{D},\mathtt{P}))_t$ ((ii), A9);
- $do(ask(\mathtt{P},\mathtt{D}))_t \to$ $(\Diamond_{t+1} do(ignore(\mathtt{D},\mathtt{P}))_{t+1} \wedge \Diamond_{t+1} do(inform(\mathtt{D},\mathtt{P}))_{t+1}$ (A8, (iv), (v), A9).

Example 3.2 [Running Example (cont'd.)] Avery suspects he has an illness, so he intends to get tested, knowing he has a right to know the results. We formalize this as the following strong belief formula:

$$RK = \Box_0 \left[do(test(\mathtt{D}))_0 \wedge do(ask(\mathtt{P},\mathtt{D})_1) \to \Box_2(\neg do(inform(\mathtt{D},\mathtt{P})_2 \to \mathsf{v}_3)\right].$$

RK should be understood as: the doctor ought to inform the patient of the test results if the patient has had the tests and has asked for his test results; otherwise, a violation occurs. That is, this is a power type of right: the duty occurs once the patient asks for the results. Note that this is not supposed to be a general definition of the power to know; it describes actions that are preconditions for the duty to hold in this setting. One is the duty-creating action of the patient, the other is a practical precondition: the patient has to get tested before he can be informed of any kind of result.

Using this formalization, we now provide the initial belief-intention database for our running example.

Example 3.3 [Running Example (cont'd.)] Initially, there are no intentions, and the only belief under consideration is: Avery has the right to know whether he is ill. Since we would like to be able to reason about obligations and what happens when a violation occurs, we use RK from Example 3.2 to formalize the right to know. Formally, the initial belief-intention database is (SB, I), where

$$SB_0 = Cn(RK) \text{ and } I_0 = \emptyset.$$

Because the patient has no action he intends to carry out, his set of weak beliefs $WB(SB_0, I_0) = SB_0$ is the set of strong beliefs.

Next, we add two intentions using the intention revision operator. Notice that we actually have two agents. Avery is the agent we consider from the planning point of view, and he reasons about the doctor's obligation when planning. He derives the doctor's obligation from his weak beliefs since he still needs to ask to be informed, and he reasons from his strong beliefs after he has made his request.

Example 3.4 [Running Example, revision with intentions (cont'd.)] After a process of planning, the following two intentions are added: $i_1 = (test(\mathtt{D}), 0)$ and $i_2 = (ask(\mathtt{P}, \mathtt{D}), 1)$. Since both these intentions cohere with the current beliefs, they can simply be added to the intention database.

More formally, using postulates [P3] and [P4] (Def. 2.7), we obtain

$$((SB_0, I_0) \otimes i_1) \otimes i_2 = (SB_1, I_1),$$

where $SB_1 = SB_0 = Cn(RK)$ (revision of intentions cannot change strong beliefs), and $I_1 = \{(test(\mathtt{D}), 0), (ask(\mathtt{P}, \mathtt{D}), 1)\}$.

Note that $WB(SB_1, I_1) \vdash \Box_2(\neg do(inform(\mathtt{D}, \mathtt{P})_2 \rightarrow \mathtt{v}_3)$ (Def. 2.4), which means that that the doctor should inform the patient of his test results at time 2; otherwise a violation occurs at time 3. We simply consider the obligation derived from weak beliefs as the result of exercising a legal power.

This is the point of power type of rights: one can plan with them with the knowledge that by carrying out these actions, the other party will have a duty. Hence, if carrying out the action (of asking) is among my intentions, then the obligation of the other person will be among those postconditions that depend on the actions I intend to carry out. That is, the obligation will be derivable from weak beliefs.

Next, we model the belief database with an action a executed at time 0. This is something that was not investigated by Van Zee *et al.* [28]. We model this simply by adding the strong belief $\Box_0 do(a)_0$, which states that some action is necessarily carried out. Intuitively, this ensures that everything that follows from executing a at time 0 is now a strong belief. So, for instance, $\Box_0 post(a)_1$ now also holds, as well as everything that follows from that.

Example 3.5 [Running Example, revision with strong beliefs (cont'd.)] Next, the doctor carries out the tests, which we model by adding the strong belief

$$(SB_1, I_1) \circ \Box_0 do(test(\mathtt{D}))_0 = (SB_2, I_2),$$

where

- $SB_2 = Cn(\{RK, \Box_0 do(test(\mathtt{D}))_0\})$;
- $I_2 = I_1 = \{(test(\mathtt{D}), 0), ask(\mathtt{P}, \mathtt{D}), 1)\}$ (adding that the strong belief did not invalidate any intentions).

We can now infer the following power relationship between the patient and the doctor: if the doctor does not provide the test result upon request, there is a violation. In other words, the doctor is obliged to provide the test results:

$$SB_2 \vdash \Box_0(do(ask(\mathtt{P}, \mathtt{D})_1 \rightarrow \Box_2(\neg do(inform(\mathtt{D}, \mathtt{P})_2) \rightarrow \mathsf{v}_3).$$

Next, the patient requests his test results:

$$(SB_2, I_2) \circ \Box_0 do(ask(\mathtt{P}, \mathtt{D}))_1 = (SB_3, I_3),$$

where

- $SB_3 = Cn(\{RK, \Box_0 do(test(D))_0, \Box_0 do(ask(\mathtt{P}, \mathtt{D}))_1\})$;
- $I_3 = I_2 = I_1 = \{(test(\mathtt{D}), 0), ask(\mathtt{P}, \mathtt{D}), 1)\}$.

We can infer the next claim-right relationship between the patient and the doctor: if the doctor does not inform the patient, there is a violation:

$$SB_3 \vdash \Box_0(\neg do(inform(\mathtt{D}, \mathtt{P})_2) \rightarrow \mathsf{v}_3).$$

We will formalize the obligation, claim-right, and legal power involved in the above examples more precisely in the next section.

3.2 Obligations and Claim-Rights in the Logic of Intentions

In deontic logic, deontic concepts such as obligation and permission are considered to be deontic variants of necessity and possibility [11]. Following this tradition, our database framework represents modalities for deontic concepts utilizing temporal modalities, taking *deontic* necessity and possibility as temporal modalities of necessity or possibility *to plan what is normative.* We will define obligation, permission and prohibition in the sense of "ought to do" [11], representing them as deontic modalities on individual actions. They are defined in the style of Anderson reduction [2].

Definition 3.6 [Obligation, Permission, and Prohibition] Given $t \in \mathbb{N}$ and $a \in \mathsf{Act}$:

- an action a that is allowed to be carried out at time t, denoted as $P(a)_t$, is defined as $\Diamond_t(do(a)_t \wedge \neg \mathsf{v}_{t+1})$;
- an action a that ought to be carried out at time t, denoted as $O(a)_t$, is defined as $\Box_t(\neg do(a)_t \rightarrow \mathsf{v}_{t+1})$;
- an action a that is prohibited from being carried out at time t, denoted as $F(a)_t$, is defined as $\Box_t(do(a)_t \rightarrow \mathsf{v}_{t+1})$.

These three deontic modalities are defined on single actions but not consecutive actions (denoted by $\overline{a}$, see Def. 2.1). For instance, if $\overline{a} = (a_1, a_2)$ with $a_1, a_2 \in$ Act, then $O(\overline{a})_3$ is not a correct expression.

In the PAL language, $P(a)_t$ means that it is possible at time t to do action a and not have a violation in the next time point. Then, $O(a)_t$ means that it must be the case that if at time t action a is not executed, there is a violation in the next time points, and $F(a)_t$ means that it must be the case that if action a is executed, there is a violation in the next time point.

Next, we extend the logic of Van Zee *et al.* with a new axiom stating that it is always possible to avoid a violation.

Definition 3.7 [Avoiding Violation Axiom] We add the following axiom to the axiomatization of Van Zee *et al.* (see [28], Section 2.3) [4]: $\Diamond_t \neg v_{t+1}$.

We now obtain the following proposition. We omit the proof since it follows straightforwardly from the definition of $O(a)_t$, $P(a)_t$ and the new axiom.

Proposition 3.8 (Obligation Implies Permission) *If we add the Avoiding Violation Axiom, $O(a)_t \rightarrow P(a)_t$ is a theorem of the logic.*

To capture claim-rights, we show how our deontic concepts can be included in the strong beliefs given a belief-intention database.

Example 3.9 [Running example, claim and privilege (cont'd.)] Recall from the previous example that $SB_3 = Cn(\{RK, do(test(\mathtt{D}))_0, do(ask(\mathtt{P}, \mathtt{D}))_1\}$ and that we could then infer the following:

$$SB_3 \vdash \Box_0(\neg do(inform(\mathtt{D}, \mathtt{P})_2) \rightarrow \mathsf{v}_3).$$

Using Def. 3.6 and Axiom A5 (Def. 2.2), it follows that an obligation is inferred:

$$SB_3 \vdash O(inform(\mathtt{D}, \mathtt{P}))_2.$$

Thus, after the patient has asked for his result at time point 1, the doctor has an obligation to inform the patient of the result at time 2. Therefore, Avery now has a claim-right that the doctor should inform him of the result.

We obtain other types of deontic concepts if we update the databases differently. For instance, assume the following strong belief formula b_1:

$$\Box_0(do(test(\mathtt{D}))_0 \rightarrow \Diamond_1((inform(\mathtt{D}, \mathtt{P}))_1 \wedge \neg \mathsf{v}_2)),$$

and suppose we update the belief-intention database, after carrying out the tests specified in the planner (Example 3.5), as follows:

$$(SB_2, I_2) \circ b_1 = (SB_2', I_2').$$

Now the following permission can be inferred:

$$SB_2' \vdash P(inform(\mathtt{D}, \mathtt{P}))_1,$$

[4] Due to space constraints, we omit the semantics here, but if we add a property to the definition of the model (see [28], Def. 6) stating that in each state there exists an action transition such that in the next time moment $\neg v$ holds, then we can straightforwardly prove that the logic remains sound and strongly complete.

which states that the doctor has a permission to inform the patient of the result, which then indicates in this belief base that the patient has the privilege of requesting that the doctor informs him of the result.

Similarly, if we add the following strong belief b_2:

$$\Diamond_0(do(test(\mathtt{D}))_0 \wedge \neg \mathsf{v}_1),$$

we have this revision of the belief-intention:

$$(SB_2, I_2) \circ b_2 = (SB_2'', I_2'').$$

Now we conclude with another permission as a strong belief in this database:

$$SB_2'' \vdash P(test(\mathtt{D}))_0.$$

So the patient has the privilege, given his strong beliefs SB_2'', of expecting the doctor to carry out the tests.

The permission and prohibition of an action cannot simply be reduced to an action obligation, as shown in the following proposition. Proposition 3.10 shows how obligation, permission, and prohibition can be connected. In particular, Proposition 3.10 (iii) and (iv) shows that a variant of the dual relation between obligation and permission exists.

Proposition 3.10 *Given $t \in \mathbb{N}$ and $a \in \mathsf{Act}$, the following propositions are theorems in our logic.*

(i) $F(a)_t \leftrightarrow \neg P(a)_t$;

(ii) $F(a)_t \rightarrow \bigvee_{b \neq a} P(b)_t$;

(iii) $P(a)_t \rightarrow \bigwedge_{b \neq a} \neg O(b)_t$;

(iv) $O(a)_t \leftrightarrow \bigwedge_{b \neq a} \neg P(b)_t$.

Note that the last part of Proposition 3.10(iv) implies that if an action is obligatory, then no other action can be permitted. In our logic, the property is a consequence of a practical interpretation of A7. It clarifies our key understanding about actions from the database perspective: if an action is executed at some time point, no other action can be performed at the same time. This leads to the conclusion that if we are obligated to do action a, we are not allowed to engage in other actions as that would prevent us from executing a. This property does not necessarily fit the understanding on norms or the law from a deontic point of view, but it fits well from a database perspective.

4 Optimality and Power

In the previous section, we formalized static deontic concepts such as obligation and permission using a violation constant. But because the coherence condition of Van Zee *et al.* does not use this information, we were not able to use it when revising with new beliefs or intentions. In this section, we propose a new condition, stronger than coherence, called “optimality”: if a belief-intention database is optimal, then it is coherent, and it avoids violation states. We show that this new coherence condition can be used to revise belief-intention databases satisfying the deontic notions we proposed in the previous section.

Definition 4.1 [Optimality] Given an intention database $I = \{(b_{t_1}, t_1), \ldots, (b_{t_n}, t_n)\}$ with $t_1 < \cdots < t_n$, let

$$Opt(I) = \Diamond_0 \bigvee_{\substack{a_t \in Act : t \notin \{t_1, \ldots, t_n\} \\ a_t = b_t : t \in \{t_1, \ldots, t_n\}}} (pre(a_{t_1}, a_{t_1+1} \ldots, a_{t_n})_{t_1} \wedge \bigwedge_{t_1 \leq i \leq t_n} (do(a_i)_i \to \neg \mathsf{v}_{t_{i+1}})).$$

For a given belief-intention database (SB, I), we say that it is optimal iff SB is consistent with $Opt(I)$, i.e., $SB \nvdash \neg Opt(I)$.

Note that the above definition requires not only that the actions intended don't lead to a new violation state but also that the other possible actions that may be carried out should act as a bridge on the path from t_1 to t_n. It ensures that no new violation can occur from t_2 to t_{n+1}. For example, $Opt(\{(a, 1), (c, 3)\})$ requires the execution of some action b at time 2 bridging a at time 1 and c at time 3 without any new violations from time 2 to time 4.

Definition 4.2 [Postulates of Optimal Revision] An *intention revision function* $\bullet$ maps a belief-intention database and an intention to a belief-intention database such that

$$(SB, I) \bullet i = (SB, I')$$

where the postulates that hold for optimality are similar to the postulates for intention revision (Def. 2.7) $(P1)$–$(P5)$, except that the condition of coherence is replaced by the condition of optimality.

In order to specify the distinction between the coherence and optimality conditions, we continue our discussion of the running example and now consider the action $ignore(\mathtt{D}, \mathtt{P})$, which means that the doctor ignores the patient's request.

Example 4.3 [Coherent Intentions vs. Optimal Intentions (Ctd.)] Recall that $SB_2 = Cn(RK) \circ \Box_0(do(test)_0)$ (Example 3.5). We consider two possible intention databases:

- $I = \{(inform(\mathtt{D}, \mathtt{P}), 2)\}$;
- $I' = \{(ignore(\mathtt{D}, \mathtt{P}), 2)\}$.

Now we have the following implications:

$$Cohere(I) = \Diamond_0 pre(inform(\mathtt{D}, \mathtt{P}))_2;$$
$$Cohere(I') = \Diamond_0 pre(ignore(\mathtt{D}, \mathtt{P}))_2.$$

Thus, both I and I' cohere with strong belief SB_2. However, only I is optimal with SB_2. The intention database I' is not optimal because there is a violated state that necessarily occurs after the intended action is executed:

$$SB_2 \vdash \Box_0(do(ignore(\mathtt{D}, \mathtt{P}))_2 \to \mathsf{v}_3).$$

This formula follows from SB_2 because, informally, it means that for all the paths in which $ignore(\mathtt{D},\mathtt{P})$ is executed at time 2, violations will occur at time 3. This is true because in each such path, with $ask(\mathtt{D},\mathtt{P})$ occurring at time 1, the doctor can ignore the patient's request for his results (recall that $post(ask(\mathtt{P},\mathtt{D})) = pre(ignore(\mathtt{D},\mathtt{P}))$ and A9).

So intention base I is optimal but intention base I' is not:

$$SB_2 \vdash Opt(I) \text{ and } SB_2 \vdash \neg Opt(I').$$

Consequently, the optimal revision of the belief-intention database (SB_2, I_0) (recall that $I_0 = \emptyset$) will not incorporate the action $ignore(\mathtt{D},\mathtt{P})$ at time 2, unlike the coherent revision:

- $(SB_2, I_0) \bullet (ignore(\mathtt{D},\mathtt{P}), 2) = (SB_2, I_0)$;
- $(SB_2, I_0) \otimes (ignore(\mathtt{D},\mathtt{P}), 2) = (SB_2, \{(ignore(\mathtt{D},\mathtt{P}), 2)\})$.

We introduced optimal revision because it prevents an artificial agent (like a robot) from remaining committed to an intended action that leads to violations and helps it to make and revise legal plans. On the other hand, violations do occur in practice, and therefore we should also allow reasoning about the dynamics of intentions (like contrary-to-duty reasoning [21]) to account for those situations. We will use a coherence condition (see Example 4.5) for this purpose.

Example 4.4 [Running example, power (cont'd.)] We know that the strong beliefs set SB_2 does not contain the following two deontic concepts:

$$SB_2 \nvdash O(inform(\mathtt{D},\mathtt{P}))_2 \text{ and } SB_2 \nvdash F(ignore(\mathtt{D},\mathtt{P}))_2.$$

Now by updating the database with intention $(ask(\mathtt{D},\mathtt{P}), 1)$, we can see that an obligation exists in the weak beliefs of the updated database $(SB_2, \{(ask(\mathtt{D},\mathtt{P}), 1)\})$:

$$WB((SB_2, \{(ask(\mathtt{D},\mathtt{P}), 1)\})) \vdash O(inform(\mathtt{D},\mathtt{P}))_2;$$
$$WB((SB_2, \{(ask(\mathtt{D},\mathtt{P}), 1)\})) \vdash F(ignore(\mathtt{D},\mathtt{P}))_2.$$

After the tests have been carried out, patient Avery has a Hohfeldian power. If Avery exercises that power by asking for the results, then he will have a claim-right that the doctor informs him of the result. If the patient does not intend to ask for the result, the doctor cannot be obliged to inform the patient.

If, instead of forming the intention $(ask(\mathtt{D},\mathtt{P}), 1)$, the planner has the action $ask(\mathtt{D},\mathtt{P})$ that is actually executed at time 1, then we obtain the revised strong beliefs set SB_3 (see Example 3.5). The same obligation and prohibition exist, but since the claim-right of Avery (and thus the corresponding duty of the doctor) was created by his request, now the obligation (and the prohibition) follows from the strong beliefs:

$$SB_3 \vdash O(inform(\mathtt{D},\mathtt{P}))_2 \text{ and } SB_3 \vdash F(ignore(\mathtt{D},\mathtt{P}))_2.$$

The question arises: what happens if the doctor ignores the request, violating her duty? This scenario leads to contrary-to-duty reasoning [21]. It is very intuitive to say that Avery's right to know his results must include a "solution" for when the newly created claim-right's corresponding duty is violated. Indeed, Avery has a new intention for which this violation is a precondition: a complaint to the hospital director. [5] The example below shows what can be done in the current version of the logic.

Example 4.5 [Running example, contrary-to-duty reasoning (cont'd.)] Assume that following Avery's request to the doctor at time 1, the doctor intends to ignore him:

$$(SB_3, I_0) \otimes (\mathit{ignore}(\mathtt{D},\mathtt{P}), 2) = (SB_3, I_4),$$

where $I_4 = \{(\mathit{ignore}(\mathtt{D},\mathtt{P}), 2)\}$. To recover from this bad situation, Avery will have a new intention: complain to the hospital director. Intuitively, this corresponds to contrary-to-duty scenarios in deontic logic literature [21], which is about how to recover when the primary obligation is violated. Therefore we have:

$$(SB_3, I_4\}) \otimes (\mathit{complain}(\mathtt{P},\mathtt{HD}), 3) = (SB_3, I_5),$$

where $I_5 = \{(\mathit{ignore}(\mathtt{D},\mathtt{P}), 2), (\mathit{complain}(\mathtt{P},\mathtt{HD}), 3)\}$. Recall that we assume that the database is shared. After the doctor understands that Avery intends to complain to the hospital director, she revises her intention and decides to let Avery know his test results:

$$(SB_3, I_5\}) \otimes (\mathit{inform}(\mathtt{D},\mathtt{P}), 2) = (SB_3, I_6).$$

Here we have that $I_6 = \{(\mathit{inform}(\mathtt{D},\mathtt{P}), 2)\}$. The intention $(\mathit{ignore}(\mathtt{D},\mathtt{P}), 2)$ is dropped because only one intention is possible at time 2, and the new intention takes priority (according to P2 from Def. 2.7). Then $(\mathit{ignore}(\mathtt{D},\mathtt{P}), 2)$ must be dropped as well because its precondition will not hold at time 3.

It would be rather intuitive to allow Avery to model conditional planning by adding to the database both his intention to complain and his intention to submit a request at time 3, depending on how the situation develops (i.e., whether the doctor informs him or ignores his request). But they have incompatible preconditions. The precondition for the complaint action is the postcondition of the ignore action, while the precondition for applying for early retirement requires that Avery is informed [6]). The agent will drop the action whose precondition is not met. In any case, the current system does not allow two intentions at the same time point, so we leave this as future work.

5 Conclusions and Future Work

Rights, including epistemic rights, influence our plans and thus the intentions we assume or discard. Avery wouldn't have gotten tested if he hadn't believed that he would get the information he needed to apply for early re-

[5] This complaint action is very similar to asking for test results; it imposes a duty on the hospital director to inform Avery (or make the doctor inform Avery). This duty can also be violated, but we do not go that far into the reasoning in this paper.

[6] For the sake of simplicity, we haven't formally added the action *submit*(P) and its pre- and postconditions to the language since we haven't used them in the example.

tirement. In order to accommodate reasoning about normative positions in a framework, we need basic deontic concepts such as obligation and permission: these could be introduced through a violation constant. We also need some formalism to express the nature of power: that some specific power action can result in changes to normative positions. We could express this by updating obligations so that weak beliefs become strong beliefs once the duty-bound action has been carried out. Additionally, one of the most characteristic features of the theory of normative positions is that we consider pairs of agents and their relations. In this paper, we considered only two agents, thus the relation between their normative positions could be handled tacitly.

We employed the PAL temporal logic of intentions [28] to reason about obligations, permissions, and rights by modeling the dynamics of intentions and beliefs. We were able to model obligations and claim-rights directly in the PAL framework and without extending it in any way. However, we did extend the revision of belief-intention databases in two ways. First, we introduced Optimal Revision, which revises the databases so that no violation can occur and prevents artificial agents from having illegal intentions. Secondly, we introduced revision of databases after actions have been carried out in order to model the nature of power (transforming weak beliefs into strong beliefs). This framework thus introduces a new way of characterizing Hohfeldian rights in practical reasoning.

In conclusion, this paper contributes to closing the gap between reasoning about rights and practical reasoning. On the one hand, the deontic concepts introduced to the framework make it possible to align the plans of artificial agents with norms. These agents are (or will be) subject to normative expectations and will have normative positions based on the deontic concepts and optimality condition involved in planning to make these possible. On the other hand, deontic logic, including the theory of normative positions, is ultimately about defining and reasoning about the normative aspect of actions. A richer action logic contributes to fulfilling its full potential.

Our future research needs to address the "ought to be" question. When it comes to rights, it seems very natural at first to talk about actions, and so "ought to do" appears to be an adequate concept to work with. In deontic logic, it is also very natural to consider "ought to be" and compare this to "ought to do" [11]. It is particularly relevant if we consider the planning aspect: the normative goal is taken as an "ought to be", and it is the role of the planner to assign the obligation to an agent to fulfill the normative plan. [7] However, from a technical point of view, defining "ought to be" in the database is more complicated than defining "ought to do". We cannot simply represent "It ought to be the case that χ at time t" as $\Box_t(\neg\chi_t \to \mathsf{v}_t)$, because axiom $\mathsf{A8}$ makes the temporal modality redundant. To maintain the temporal necessity of

[7] In fact, this also fits what happens with rights. For instance, a legislative agent that signs the Convention of Human Rights is obliged to assign corresponding duties in its own legal system.

planning normatively while avoiding temporal redundancy, one could consider the following definition of "ought to be": $O(\chi_t) := \Box_{t-1}(\neg\chi_t \to \mathsf{v}_t)$. This states: "It is necessary to plan at time $t-1$ that χ will be the case at time t if no violation occurs", which makes sense as a way to describe χ as a normative goal for the planner. However, the proper formalization of "ought to be" remains to be studied.

As compared to existing theories of agents and norms, our proposal highlights the crucial role of belief and intention in normative reasoning. Traditional logic-based methods, including dynamic deontic logic [20], see-to-it-that (STIT) logics [3], and labeled transition systems [25], encompass a wide range of deontic and temporal operators, which are interpreted using semantic models like CTL*. Our logic also uses CTL*-like models and fairly simple syntax based on PAL [28]. The framework is expressive enough to model rights and define deontic operators but is simple enough to perform AGM-style revision of belief and intention, and is therefore suitable for practical reasoning. It can be extended to address issues related to physical or normative constraints, such as environmental persistence [25], multi-agent interaction within the context of personal intentions [4,10], and the trade-off between violation and compliance [6]. We leave these topics for future work.

Acknowledgments

We would like to express our sincere gratitude for the valuable comments provided by the three referees of DEON 2023. This work was supported by the *Fonds National de la Recherche* of Luxembourg through the project Deontic Logic for Epistemic Rights (OPEN O20/14776480) and the project INTEGRAUTO (INTER/AUDACE/21/16695098). L. van der Torre is financially supported by the CHIST-ERA grant CHIST-ERA19-XAI (G.A. INTER/CHIST/19/14589586) and by the European Union's Justice program under grant agreement 101007420 (ADELE). Huimin Dong, adjunct research fellow at the Institute of Logic and Cognition at Sun Yat-sen University, is supported by the Chinese Ministry of Education (MOE) project of the Key Research Institute of Humanities and Social Sciences at Universities (22JJD720021).

References

[1] Alchourrón, C. E., P. Gärdenfors and D. Makinson, *On the logic of theory change: Partial meet contraction and revision functions*, Journal of Symbolic Logic **50** (1985), pp. 510–530.

[2] Anderson, A. R., *A reduction of deontic logic to alethic modal logic*, Mind **67** (1958), pp. 100–103.
URL `http://www.jstor.org/stable/2251344`

[3] Belnap, N., M. Perloff and M. Xu, "Facing the future: agents and choices in our indeterminist world," Oxford University Press, 2001.

[4] Bratman, M., "Intention, plans, and practical reason," Harvard University Press, 1987.

[5] Broersen, J., M. Dastani, J. Hulstijn, Z. Huang and L. van der Torre, *The boid architecture: Conflicts between beliefs, obligations, intentions and desires*, in: *Proceedings*

of the Fifth International Conference on Autonomous Agents, AGENTS '01 (2001), p. 9–16.
URL `https://doi.org/10.1145/375735.375766`

[6] Broersen, J. and L. van der Torre, *Ten problems of deontic logic and normative reasoning in computer science*, Lectures on Logic and Computation: ESSLLI 2010 Copenhagen, Denmark, August 2010, ESSLLI 2011, Ljubljana, Slovenia, August 2011, Selected Lecture Notes (2012), pp. 55–88.

[7] Chellas, B. F., "Modal Logic. An Introduction." Cambridge University Press, 1980.

[8] Cohen, P. R. and H. J. Levesque, *Intention is choice with commitment*, Artificial intelligence **42** (1990), pp. 213–261.

[9] Dong, H. and O. Roy, *Dynamic logic of legal competences*, Journal of Logic, Language and Information **30** (2021), pp. 701–724.

[10] Gilbert, M., *Shared intention and personal intentions*, Philosophical studies **144** (2009), pp. 167–187.

[11] Hilpinen, R. and P. McNamara, *Deontic logic: A historical survey and introduction*, Handbook of Deontic Logic and Normative Systems **1** (2013), pp. 3–136.

[12] Jones, A. J. I. and M. Sergot, *A Formal Characterisation of Institutionalised Power*, Logic Journal of the IGPL **4** (1996), pp. 427–443.

[13] Kanger, S. and H. Kanger, *Rights and parliamentarism*, Theoria **32** (1966), pp. 85–115.

[14] Li, X., D. Gabbay and R. Markovich, *Dynamic Deontic Logic for Permitted Announcements*, in: *Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning*, 2022, pp. 226–235.

[15] Lindahl, L., "Position and change: A study in law and logic," Synthese Library, Springer Dordrecht, 1977, 1 edition.

[16] Makinson, D., *On the formal representation of rights relations*, Journal of Philosophical Logic **15** (1986), pp. 403–425.

[17] Markovich, R., *Understanding Hohfeld and Formalizing Legal Rights: the Hohfeldian Conceptions and Their Conditional Consequences*, Studia Logica **108** (2020).
URL `onlinefirst:https://doi.org/10.1007/s11225-019-09870-5`

[18] Markovich, R. and O. Roy, *Cause of action and the right to know. a formal conceptual analysis of the texas senate bill 25 case*, in: *Legal Knowledge and Information Systems - JURIX 2021*, Frontiers in Artificial Intelligence and Applications **346**, IOS Press, 2021 pp. 217–224.

[19] Markovich, R. and O. Roy, *Formalizing the right to know: Epistemic rights as normative positions*, in: *The First International Workshop on Logics for New-Generation Artificial Intelligence (LNGAI 2021)*, 2021, pp. 154–159.

[20] Meyer, J.-J. C. et al., *A different approach to deontic logic: deontic logic viewed as a variant of dynamic logic.*, Notre Dame J. Formal Log. **29** (1988), pp. 109–136.

[21] Prakken, H. and M. Sergot, *Contrary-to-duty obligations*, Studia Logica **57** (1996), pp. 91–115.

[22] Rao, A. S. and M. P. Georgeff, *Modeling rational agents within a bdi-architecture*, Readings in agents , pp. 317–328.

[23] Reynolds, M., *An axiomatization of full computation tree logic*, The Journal of Symbolic Logic **66** (2001), pp. 1011–1057.

[24] Sergot, M., *Normative Positions*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, College Publications, 2013 pp. 353–406.

[25] Sergot, M. and R. Craven, *The deontic component of action language nc+*, in: *Deontic Logic and Artificial Normative Systems: 8th International Workshop on Deontic Logic in Computer Science, DEON 2006, Utrecht, The Netherlands, July 12-14, 2006. Proceedings 8*, Springer, 2006, pp. 222–237.

[26] Shoham, Y., *Logical theories of intention and the database perspective*, Journal of Philosophical Logic **38** (2009).

[27] Thomason, R. H., *Practical reasoning: Problems and prospects*, in: X. P. R. v. d. M. Dov Gabbay, John Horty and L. van der Torre, editors, *Handbook of Deontic Logic Vol. 2*, College Publications, 2021 pp. 463–498.

[28] van Zee, M., D. Doder, L. van der Torre, M. Dastani, T. Icard and E. Pacuit, *Intention as commitment toward time*, Artificial Intelligence **283** (2020), p. 103270.
URL `https://www.sciencedirect.com/science/article/pii/S0004370220300308`

[29] Watson, L., *The right to know: Epistemic rights, and why we need them*, Australasian Journal of Philosophy **100** (2022), pp. 426–427.

A proof theory for admissibility in two-player (deontic) games

Edi Pavlović [1]

Munich Center for Mathematical Philosophy (MCMP)
Geschwister-Scholl-Platz 1, D-80539 München

Frederik Van De Putte [2]

Erasmus University Rotterdam
Burgemeester Oudlaan 50, 3062 PA, Rotterdam

Abstract

We study a **STIT** logic for two agents, augmented with three deontic constants $1, 1_i, 1_j$. The constants express, respectively, admissibility for the group $\{i, j\}$, admissibility for agent i, and admissibility for agent j. Our semantics uses deontic game models [25,5], where an action X is admissible for an (individual or group) agent if and only if X is not weakly dominated by any other action X' that is available to that same agent. After presenting the formal language and game-theoretic semantics, we first spell out a corresponding Kripke-semantics for the same logic. On the basis of the latter, we provide a proof system via a geometric extension of a labelled sequent calculus for **STIT** in [16]. We demonstrate the structural properties of said calculus, including height-preserving admissibility of contraction, as well as cut elimination, and then establish soundness and completeness. Finally, we illustrate the general applicability of the calculus by discussing a number of possible variations and extensions.

Keywords: Deontic logic, STIT, game models, relational models, sequent calculus, soundness and completeness

1 Introduction

The marriage between logics of agency and deontic logic has been a long and fruitful one, dating back to the famous Meinong-Chisholm thesis [8] and the work on normative positions [10,11]. With the advent of STIT logic [3], the logic of agency received renewed interest in the deontic logic community, with

[1] edi.pavlovic@lmu.de. Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 459928802. Funded by the German Research Foundation (DFG) - Project number 459928802.

[2] vandeputte@esphil.eur.nl. Funded by VIDI grant no. VI.Vidi.191.105 from the Dutch Research Council (NWO).

[9] as a key reference. What distinguishes Horty's approach from earlier work is that he uses decision-theoretic concepts and models to provide a semantics for expressions of the type "agent i ought to see to it that A" (O_iA). On Horty's *dominance act utilitarian* semantics, the latter statement is true if and only if all admissible actions of i are such that they guarantee A. Admissibility is in turn spelled out in line with the standard concept of (weak) dominance from decision theory, by quantifying over all possible combinations of actions of all other agents, and relative to an agent-independent (normative) comparison of outcomes. [3]

More recently, Tamminga and Hindriks [26] have developed an alternative, game-theoretic semantics for deontic STIT logic that abstracts from the branching time framework of traditional STIT logic and takes the agent's actions as primitive. [4] In [25,5], this semantics is used to interpret a more expressive language where the deontic operators are replaced with deontic constants 1_i expressing that "agent i is performing an admissible action". As shown in [25], O_iA can then be defined as $\mathcal{S}(1_i \to A)$, i.e., "it is settled true that if i performs an admissible action, then A", following the well-known Andersonian-Kangerian reduction of deontic logic [10,1].

While there has been a recent uptick in the proof-theoretic investigations of both STIT logic [21,31,12,16,2,17] and deontic logic [22,32,7], none of these tackle admissibility constants. [5] Proof-theoretic methods allow for a streamlined integration of semantics within a system of syntactic rules, in a manner that is highly modular. That is to say, extensions of the system do not require us to backtrack and re-check any of the previous results, but rather one needs only to establish that those still hold for the new additions. Moreover, if one considers rules of the geometric form [18,20], as is the case in the present paper, it is only necessary to ascertain that they are of correct *syntactic* format, and no further proof is required. This is especially useful when combining two considerably complex systems, as is the case here.

Desired properties enable or facilitate a wide range of insights, both philosophical and technical. Among those, invertibility of the rules (whereby, if a conclusion of a rule is derivable so are its premises) means that a counterexample to a premise is also a counterexample to the conclusion. Consequently, one can generate countermodels from a failed proof search. This allows for straightforward proofs of completeness and is utilized in this paper.

A more general property of a system is analyticity, whereby features of a derivation, such as its length or the number and weight of formulas therein, can be gleaned simply by observing the conclusion. This property is achieved by,

[3] See in particular [9, Chapter 4].

[4] Similar models are used in [24] to analyse collective obligations and team plans.

[5] Murakami [13] presents an axiomatization of one of Horty's logics, for a language that (only) features individual agency operators $\Box^i$ and individual obligation operators O_i. As shown there, the axiomatization requires no specific interactions between the deontic operators; all theorems can be derived from the STIT fragment and the way each individual obligation operator O_i interacts with the corresponding agency operator $\Box^i$.

among others, admissibility of contraction, cut, and the resultant subformula property (where only syntactic subformulas of those in the conclusion occur in any derivation). As a consequence, by merely observing the (purported) conclusion, it is possible to infer facts about its acceptability or extract more precise and substantial sub-derivations, as was recently done for deontic logic without STIT in [7]. Finally, proof theoretic characterizations are an important step towards syntactic proofs of decidability [22,12,16] and complexity results in general.

Here we make some initial headway in proof-theoretic investigation of deontic STIT logic with admissibility constants. We focus on a STIT logic for two agents with deontic constants $1, 1_i, 1_j$ that express, respectively, admissibility for the group $\{i, j\}$, admissibility for agent i, and admissibility for agent j. This represents the smallest multi-agent case. The extension to n agents, which we discuss and outline in Section 6, is not difficult, but the present case is sufficiently illustrative while offering far greater transparency. We spell out the formal language and game-theoretic semantics of the resulting logic $\mathbf{STIT^d_2}$ (in the style of [25]) in Section 2. Next, we provide a corresponding Kripke-semantics (Section 3). The latter in turn allows us to build a proof system for $\mathbf{STIT^d_2}$, via a geometric extension of a labelled sequent calculus for STIT in [16], and demonstrate its structural properties (Section 4). In Section 5 we establish soundness and completeness. Finally, we argue that our calculus can be generalized to other deontic STIT logics and logics of weak dominance in normal form games (Section 6).

2 Language and semantics of $\mathbf{STIT^d_2}$

In this section we present the formal language and game-theoretic semantics of $\mathbf{STIT^d_2}$. Our exposition mostly follows [25,5]; where we deviate from the modeling choices in this work, we explicitly mention this.

2.1 Formal language

Fix a countable set $\mathfrak{P}$ of propositional variables and two distinct agents i, j. In what follows, we use α as a metavariable for both and we let $\bar{i} = j$, and $\bar{j} = i$. The formal language $\mathfrak{L}$ is given by the following Backus-Naur form:

$$A ::= \bot \mid p \mid 1 \mid 1_\alpha \mid 0 \mid 0_\alpha \mid \neg A \mid A \supset A \mid \Box^\alpha A \mid \Diamond^\alpha A \mid \mathcal{S}A \mid \mathcal{P}A$$

where $p \in \mathfrak{P}$ and $\alpha \in \{i, j\}$. Table 1 provides an overview of the different interpretations of the non-standard constants and operators in this language. Note that we treat more symbols as primitive than is usual in a classical modal logic setting; this allows us to draw on established proof theoretic techniques to obtain the sequent calculus for the logic [18,20,15,16].

2.2 Two-agent deontic game models

Deontic game models (for the set of agents $\{i, j\}$) can be seen as normal game forms equipped with (i) a deontic assignment that tells us which combinations of actions by i and j are acceptable, and (ii) a valuation function that tells us

1	"The current state is deontically acceptable." "The current combination of actions is admissible for $\{i, j\}$."
1_i	"i performs a deontically admissible action." "i's choice is deontically admissible."
$\Box^i p$	"i sees to it that p." "i's choice guarantees that p is the case."
$\mathcal{S}p$	"p is true, no matter what i or j do."

Table 1
Natural language interpretations of some expressions in $\mathfrak{L}$.

which propositions are true given any such combination. Formally:

Definition 2.1 (Deontic Game Model) *A* (two-agent) deontic game model *M is a quadruple $\langle \mathsf{A}_i, \mathsf{A}_j, d, V \rangle$, where A_i and A_j are non-empty sets of actions available to agent i, resp. j, where $d : \mathsf{A}_i \times \mathsf{A}_j \to \{0, 1\}$ is a deontic value assignment such that there is at least one a in $\mathsf{A}_i \times \mathsf{A}_j$ with $d(a) = 1$, and $V : \mathfrak{P} \to \wp(\mathsf{A}_i \times \mathsf{A}_j)$ is a valuation function.*

In what follows, we use A to abbreviate $\mathsf{A}_i \times \mathsf{A}_j$, where the deontic game model in question is clear from the context. A is called the set of *action profiles* of the given deontic game model.

Note that, in contrast to [25,5], we do not presuppose that A_i and A_j are finite. Such an assumption would arguably render the logic non-compact. [6] As a consequence of lifting the finiteness condition, the simple dominance relation on actions (cf. Definition 2.2) is also no longer smooth, which does have an impact on the logic in question — we return to this point in Section 6.

To illustrate the above definition, consider the model M_1 depicted in Figure 1. Here, rows represent actions of i and columns represent actions of j. Each cell in the diagram thus corresponds to an action profile. The deontic value assignment and the valuation function are represented by the 1s and 0s, and by the propositional variables that occur in these cells. Thus, for instance, $d(c_i, b_j) = 0$ and $V(p) = \{(a_i, a_j), (c_i, c_j)\}$.

	a_j	b_j	c_j
a_i	$0/p$	1	1
b_i	$1/q$	$1/q$	$0/q$
c_i	1	0	$0/p$

Fig. 1. Deontic game model M_1

Given any such model, formulas are evaluated relative to action profiles. For instance, in the model M_1 given above, $\Box^i q$ is true at the action profile

[6] In particular, one can construct an infinite set that expresses that there are infinitely many distinct action profiles: $\Gamma = \{\mathcal{P}(\neg p_1 \wedge \ldots \wedge \neg p_n \wedge p_{n+1}) \mid n \geq 0\}$. Every finite subset of Γ can be satisfied by some finite deontic game model, but Γ cannot. If one imposes the stronger condition that $|A_i \times A_j| \leq k$ for some $k \in \mathbb{N}$, then the logic is again compact.

(b_i, a_j), since given b_i, it is guaranteed that q is the case. In contrast, $\Box^i q$ is false at (a_i, a_j), since a_i does not in itself guarantee that q is the case.

Similarly, whether α performs an admissible action depends on the action profile in question. To specify truth-conditions for 1_α, we compare the actions in A_α by way of a *simple dominance* relation $\succeq^\alpha_M$:[7]

Definition 2.2 (Simple Dominance) *Let* $M = \langle \mathsf{A}_i, \mathsf{A}_j, d, V\rangle$ *be a deontic game model and let* $\alpha \in \{i, j\}$. *Then*

$$a_\alpha \succeq^\alpha_M b_\alpha \quad \textit{iff} \quad \textit{for all } c_{\overline{\alpha}} \in \mathsf{A}_{\overline{\alpha}} \textit{ it holds that } d(a_\alpha, c_{\overline{\alpha}}) \geq d(b_\alpha, c_{\overline{\alpha}}).$$

Weak dominance is the strict counterpart of simple dominance: $a_\alpha \succ^\alpha_M b_\alpha$ if and only if $a_\alpha \succeq^\alpha_M b_\alpha$ and $b_\alpha \not\succeq^\alpha_M a_\alpha$. Finally, an action is *deontically admissible* in deontic game model M if it is $\succeq^\alpha_M$-maximal:

Definition 2.3 (Deontic Admissibility) *Let* $M = \langle \mathsf{A}_i, \mathsf{A}_j, d, V\rangle$ *be a deontic game model and* $\alpha \in \{i, j\}$. *Then the set of* α*'s* deontically admissible actions *in* M *is given by*

$$\mathsf{Adm}_\alpha(M) \quad = \quad \{a_\alpha \in \mathsf{A}_\alpha : \textit{there is no } b_\alpha \in \mathsf{A}_\alpha \textit{ such that } b_\alpha \succ^\alpha_M a_\alpha\}.$$

So for instance, $\mathsf{Adm}_i(M_1) = \{a_i, b_i\}$ and $\mathsf{Adm}_j(M_1) = \{a_j, b_j\}$.

We are now in a position to state the truth-conditions for $\mathbf{STIT^d_2}$:[8]

Definition 2.4 (Truth-Conditions) *Where* $M = \langle \mathsf{A}_i, \mathsf{A}_j, d, V\rangle$ *is a deontic game model,* $a \in \mathsf{A}$, $\alpha \in \{i, j\}$, *and* $p \in \mathfrak{P}$*:*

$$\begin{array}{lll}
M, a \models p & \textit{iff} & a \in V(p) \\
M, a \models 1 & \textit{iff} & d(a) = 1 \\
M, a \models 0 & \textit{iff} & d(a) \neq 1 \\
M, a \models 1_\alpha & \textit{iff} & a_\alpha \in \mathsf{Adm}_\alpha(M) \\
M, a \models 0_\alpha & \textit{iff} & a_\alpha \notin \mathsf{Adm}_\alpha(M) \\
M, a \models \mathcal{S}A & \textit{iff} & \textit{for all } b \in \mathsf{A} \textit{ it holds that } M, b \models A \\
M, a \models \mathcal{P}A & \textit{iff} & \textit{for some } b \in \mathsf{A} \textit{ it holds that } M, b \models A \\
M, a \models \Box^\alpha A & \textit{iff} & \textit{for all } b \in \mathsf{A} \textit{ with } b_\alpha = a_\alpha \textit{ it holds that } M, b \models A \\
M, a \models \Diamond^\alpha A & \textit{iff} & \textit{for some } b \in \mathsf{A} \textit{ with } b_\alpha = a_\alpha \textit{ it holds that } M, b \models A.
\end{array}$$

3 Kripke semantics for $\mathbf{STIT^d_2}$

We now provide a Kripke-semantics for $\mathbf{STIT^d_2}$ as an intermediary between game models and sequent calculi, which will allow us to draw on well-known techniques for the proof theoretic characterization of **STIT**-logics.[9] We do so in two steps: first we define a more general class of models for $\mathfrak{L}$, and next we

[7] Here, we follow terminology recently introduced in [6] and deviate from [25,5], in order to be more in line with common terminology in decision and game theory. Horty [9] uses “strong dominance” for what is called weak dominance here.

[8] Here and below, we omit the standard truth-conditions for the classical connectives.

[9] The link between normal game forms and (Kripke-semantics for) normal modal logics of type **S5** is meanwhile well-documented. Some key references are [30,29,28], cf. also [27, Section 2.6] for an introduction to this area.

impose additional conditions on them that ensure that the deontic constants 1_α and 0_α get their intended meaning.

Definition 3.1 (Relational quasi-model) *A* relational quasi-model *is a quintuple $M = \langle W, \sim^i, \sim^j, d, V\rangle$, where W is non-empty, $\sim^i$ and $\sim^j$ are equivalence relations over W, $d : W \to \{0,1\}$ is a deontic value assignment, and $V : \mathfrak{P} \cup \{1_i, 0_i, 1_j, 0_j\} \to \wp(W)$ is a valuation function, and such that each of the following conditions hold:*

- Independence of Agents *(IOA): for all $w, w' \in W$, there is a $w^* \in W$ such that $w \sim_i w^*$ and $w' \sim_j w^*$*
- Determinism *(Det): for all $w, w' \in W$, if $w \sim^i w'$ and $w \sim^j w'$, then $w = w'$*
- Deontic Consistency *(D): there is some $w \in W$ such that $d(w) = 1$*

Definition 3.2 (Valuation) *Where $M = \langle W, \sim^i, \sim^j, d, V\rangle$ is a relational quasi-model, $A, B \in \mathfrak{L}$, and $w \in W$:*

$$\begin{array}{lll}
M, w \models A & \textit{iff} & a \in V(A) \textit{ for } A \in \mathfrak{P} \cup \{1_i, 0_i, 1_j, 0_j\} \\
M, w \models 1 & \textit{iff} & d(w) = 1 \\
M, w \models 0 & \textit{iff} & d(w) \neq 1 \\
M, w \models \mathcal{S}A & \textit{iff} & \textit{for all } w' \in W, \textit{ it holds that } M, w \models A \\
M, w \models \mathcal{P}A & \textit{iff} & \textit{for some } w' \in W,\ M, w \models A \\
M, w \models \Box^\alpha A & \textit{iff} & \textit{for all } w' \in W \textit{ such that } w \sim^\alpha w', \textit{ it holds that } M, w' \models A \\
M, w \models \Diamond^\alpha A & \textit{iff} & \textit{for some } w' \in W \textit{ such that } w \sim^\alpha w', \textit{ it holds that } M, w' \models A
\end{array}$$

Henceforth, let $|w|_\alpha = \{w' \in W \mid w \sim^\alpha w'\}$ and let $\mathsf{A}_\alpha(M) = \{|w|_\alpha \mid w \in W\}$. It can be easily observed that, save for the deontic constants 1_α and 0_α, there is a one-to-one mapping from deontic game models to relational quasi-models and vice versa that preserves equivalence. In particular, actions of an agent α in a deontic game model correspond to equivalence classes $|w|_\alpha$ in the corresponding relational quasi-model, and action profiles correspond to worlds. The conditions (IOA) and (Det) ensure that every combination of a given action $X \in \mathsf{A}_i(M)$ with an action $Y \in \mathsf{A}_j(M)$ coincides with a unique world w, i.e. $X \cap Y = \{w\}$, and hence with a unique action profile in the corresponding deontic game model.[10]

In order to obtain full equivalence of the semantics, we need to ensure that the deontic constants 1_α (0_α) are true (false) in exactly those states w for which $|w|_\alpha$ is "deontically admissible". We first define the latter notion for relational quasi-models and then introduce the relevant conditions. In what follows, we extend the deontic function d so that it ranges over singleton sets of worlds, putting $d(\{w\}) = d(w)$.

Definition 3.3 *Let $X, X' \in \mathsf{A}_\alpha(M)$. Then $X \succeq^\alpha_M X'$ iff for all $Y \in \mathsf{A}_{\overline{\alpha}}(M)$, $d(X \cap Y) \geq d(X' \cap Y)$. $X \succ^\alpha_M X'$ iff $X \succeq^\alpha_M X'$ and $X' \not\succeq^\alpha_M X$. $X \in \mathsf{A}_\alpha(M)$ is*

[10] See [33] for an investigation of non-deterministic STIT-models and their relation to deterministic ones.

deontically admissible *for α in M ($X \in \mathsf{Adm}_\alpha(M)$) iff there is no $X' \in \mathsf{A}_\alpha(M)$ such that $X' \succ_M^\alpha X$.*

Definition 3.4 (Relational model) *$M = \langle W, \sim^i, \sim^j, d, V\rangle$ is a* relational $\mathbf{STIT_2^d}$-model *iff M is a relational quasi-model and each of the following hold for $\alpha \in \{i, j\}$ and for all $w \in W$:*

$w \in V(1_\alpha)$ iff $|w|_\alpha \in \mathsf{Adm}_\alpha(M)$

$w \in V(0_\alpha)$ iff $|w|_\alpha \notin \mathsf{Adm}_\alpha(M)$.

To avoid redundancy we indicate $|w|_\alpha \succeq_M^\alpha |w'|_\alpha$ by $w \succeq_M^\alpha w'$, and further omit an explicit reference to M, which we call simply 'relational model', when clear from context. The presentation so far allows a simple mapping of relational and game models. To likewise facilitate the mapping to sequent calculi, we will characterize the above two conditions in the form closer to an implication:

Lemma 3.5 *M is a relational $\mathbf{STIT_2^d}$-model iff M is a quasi-model and each of the following hold, for all $X \in \mathsf{A}_\alpha(M)$:*
(BA$_\alpha$) Either $X \subseteq V(1_\alpha)$ or $X \subseteq V(0_\alpha)$, but not both.
(BND$_\alpha$) If $X \subseteq V(1_\alpha)$, then for all $X' \in \mathsf{A}_\alpha(M)$: if $X' \succeq_M^\alpha X$ then $X \succeq_M^\alpha X'$.
(NBD$_\alpha$) If $X \subseteq V(0_\alpha)$, then there is an $X' \in \mathsf{A}_\alpha(M)$ s.t. $X' \succeq_M^\alpha X$ and $X \not\succeq_M^\alpha X'$.

Proof. (L-R) This follows immediately from Definitions 3.3 and 3.4.

(R-L) By (BA$_\alpha$), for every w: $w \in V(1_\alpha)$ or $w \in V(0_\alpha)$ (but not both). So it suffices to prove only one of the two equivalences in Definition 3.4. Suppose first that $w \in V(0_\alpha)$. By (BA$_\alpha$), $|w|_\alpha \subseteq V(0_\alpha)$. By (NBD$_\alpha$) there is a $Y \in \mathsf{A}_\alpha(M)$ such that $Y \succeq_M^\alpha |w|_\alpha$. Hence $|w|_\alpha \notin \mathsf{Adm}_\alpha(M)$. The reasoning for the other direction is analogous, using (BND$_\alpha$). □

Intuitively, these conditions express the facts that (BA$_\alpha$): optimality (being the **b**est) belongs to **a**ctions, (BND$_\alpha$): **b**est actions are **n**ot (strictly) **d**ominated, and (NBD$_\alpha$): if an action is **n**ot the **b**est, then it is **d**ominated by another action.

In the following section we formulate a sequent calculus for this logic by unraveling the conditions into rules in a geometric format, enabling us to briefly and schematically demonstrate all the prerequisite structural rules. This will be made possible by representing actions via their elements, e.g. the condition (BA$_\alpha$) is represented via a geometric implication $(w : 1_\alpha \wedge w \sim^\alpha w') \rightarrow w' : 1_\alpha$, mirroring the corresponding step in the proof of Lemma 3.5 above.

4 Sequent calculi for $\mathbf{STIT}_2^d$

The basic unit of sequent calculus is a *sequent*, of the form $\Gamma \Rightarrow \Delta$, where Γ, Δ are multisets of formulas. All the rules of sequent calculi then consist of one sequent, written below the inference line, which is its *conclusion*, and one or more sequents above the line called its *premises.* All the formulas except Γ and Δ are called *active* formulas of the rule if they occur in the premise(s) and *principal* if they occur in the conclusion of the rule. Γ and Δ are called a *context* of the rule. A *branch* is a series of sequents, starting with the endsequent, in which every element is a conclusion of a rule that the following element is a premise of (two-premise rules thus split the branches).

The *height* of a derivation is the length (number of consecutive applications of derivation rules) of its longest branch. Derivability with height bounded by n is then indicated by $\vdash_n$, and height-preservation (hp) means that in the resulting derivation height is not increased. When a rule and a semantic constraint bear the same name, they can be distinguished by the latter appearing in parentheses.

Initial sequents: $w : p, \Gamma \Rightarrow \Delta, w : p$ $\qquad$ $w : \bot, \Gamma \Rightarrow \Delta$

Propositional rules: Standard G3cp, negation and implication only.

Modal rules:

$$\frac{w' \sim^{\alpha} w, w : \Box^{\alpha} A, w' : A, \Gamma \Rightarrow \Delta}{w' \sim^{\alpha} w, w : \Box^{\alpha} A, \Gamma \Rightarrow \Delta}\,\mathrm{L}\Box^{\alpha} \qquad \frac{w' \sim^{\alpha} w, \Gamma \Rightarrow \Delta, w' : A}{\Gamma \Rightarrow \Delta, w : \Box^{\alpha} A}\,R\Box^{\alpha}$$

$$\frac{w \sim^{\alpha} w', w' : A, \Gamma \Rightarrow \Delta}{w : \Diamond^{\alpha} A, \Gamma \Rightarrow \Delta}\,\mathrm{L}\Diamond^{\alpha} \qquad \frac{w \sim^{\alpha} w', \Gamma \Rightarrow \Delta, w : \Diamond^{\alpha} A, w' : A}{w \sim^{\alpha} w', \Gamma \Rightarrow \Delta, h : \Diamond^{\alpha} A}\,\mathrm{R}\Diamond^{\alpha}$$

$$\frac{w' : A, w : \mathcal{S}A, \Gamma \Rightarrow \Delta}{w : \mathcal{S}A, \Gamma \Rightarrow \Delta}\,\mathrm{L}\mathcal{S} \qquad \frac{\Gamma \Rightarrow \Delta, w' : A}{\Gamma \Rightarrow \Delta, w : \mathcal{S}A}\,\mathrm{R}\mathcal{S}$$

$$\frac{w' : A, \Gamma \Rightarrow \Delta}{w : \mathcal{P}A, \Gamma \Rightarrow \Delta}\,\mathrm{L}\mathcal{P} \qquad \frac{\Gamma \Rightarrow \Delta, w : \mathcal{P}A, w' : A}{\Gamma \Rightarrow \Delta, w : \mathcal{P}A}\,\mathrm{R}\mathcal{P}$$

Rules for relational atoms:

$$\frac{w = w, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta}\,\mathit{Refl}_{=} \qquad \frac{At(w'), w' = w, At(w), \Gamma \Rightarrow \Delta}{w = w', At(w), \Gamma \Rightarrow \Delta}\,\mathit{Repl}_{=}$$

$$\frac{w \sim^{\alpha} w', w \sim^{\overline{\alpha}} w', w = w', \Gamma \Rightarrow \Delta}{w \sim^{\alpha} w', w \sim^{\overline{\alpha}} w', \Gamma \Rightarrow \Delta}\,\mathit{Det} \qquad \frac{w \sim^{\alpha} w, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta}\,\mathit{Refl}_{\sim^{\alpha}}$$

$$\frac{w_2 \sim^{\alpha} w_3, w_1 \sim^{\alpha} w_2, w_1 \sim^{\alpha} w_3, \Gamma \Rightarrow \Delta}{w_1 \sim^{\alpha} w_2, w_1 \sim^{\alpha} w_3, \Gamma \Rightarrow \Delta}\,\mathit{Etrans}_{\sim^{\alpha}}$$

$$\frac{w_1 \sim^{\alpha} w' \sim^{\overline{\alpha}} w_2, w_1 \sim^{\alpha} w'_1, w_2 \sim^{\overline{\alpha}} w'_2, \Gamma \Rightarrow \Delta}{w_1 \sim^{\alpha} w'_1, w_2 \sim^{\overline{\alpha}} w'_2, \Gamma \Rightarrow \Delta}\,\mathit{Ind}$$

- w' is fresh (the *eigenvariable*) in $\mathrm{R}\Box^{\alpha}$, $\mathrm{L}\Diamond^{\alpha}$, $\mathrm{R}\mathcal{S}$, $\mathrm{L}\mathcal{P}$ and *Ind*. $At(w)$ is either a relational atom containing w, or an atomic formula labelled by w.

Fig. 2. $\mathrm{G3STIT}_2$

The base for our system is (a slight modification of) G3STIT [16], a *labelled* sequent calculus [20], whereby for a countable set of labels $\mathfrak{H}$, every formula $A \in \mathfrak{L}$ combines with a label $w \in \mathfrak{H}$ to produce a *labelled formula* $w : A$. Simplified to two agents and using worlds instead of moment/history pairs, this base is introduced in Figure 2.[11] Displayed formulas which are not labelled are referred to as *relational atoms*.

As previously discussed, the language $\mathfrak{L}$ extends the base language of STIT logic with three deontic constants. Henceforth, let subscript ϵ stand for either i, j, or an empty string of symbols. The sequent calculus G3STIT$_2^d$ is obtained by extending G3STIT$_2$ with the deontic rules in Figure 3.[12]

Notice that the rules 01_ϵ and $\overline{01_\epsilon}$ state that each pair of constants is mutually exclusive and jointly exhaustive, rules for $\succeq$ express its inferential behavior per Definition 2.2, rule PB expresses the condition (D) of Definition 3.1, while the remaining rules capture the conditions from Lemma 3.5 as already discussed after that Lemma.

Deontic part:

$$\frac{w\!:\!1_\epsilon, \Gamma \Rightarrow \Delta \qquad w\!:\!0_\epsilon, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta}\;01_\epsilon \qquad \frac{}{w\!:\!1_\epsilon, w\!:\!0_\epsilon, \Gamma \Rightarrow \Delta}\;\overline{01_\epsilon}$$

$$\frac{w'\!:\!1, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta}\;\text{PB} \qquad \frac{w \sim^\alpha w', w \succeq^\alpha w'', w' \succeq_\alpha w'', \Gamma \Rightarrow \Delta}{w \sim^\alpha w', w \succeq^\alpha w'', \Gamma \Rightarrow \Delta}\;\succeq^\alpha_L$$

$$\frac{w'\!:\!1_\alpha, w\!:\!1_\alpha, w \sim^\alpha w', \Gamma \Rightarrow \Delta}{w\!:\!1_\alpha, w \sim^\alpha w', \Gamma \Rightarrow \Delta}\;\text{BA}_\alpha \qquad \frac{w \sim^\alpha w', w \succeq^\alpha w'', w \succeq_\alpha w', \Gamma \Rightarrow \Delta}{w'' \sim^\alpha w', w \succeq^\alpha w'', \Gamma \Rightarrow \Delta}\;\succeq^\alpha_R$$

$$\frac{w \succeq^\alpha w', w \sim^{\overline{\alpha}} w', w'\!:\!1, w\!:\!1, \Gamma \Rightarrow \Delta}{w \succeq^\alpha w', w \sim^{\overline{\alpha}} w', w'\!:\!1, \Gamma \Rightarrow \Delta}\;\succeq^\alpha$$

$$\frac{w \succeq^\alpha w_1, w \sim^{\overline{\alpha}} w_1, w\!:\!1_\alpha, \Gamma \Rightarrow \Delta \qquad w_1 \sim^\alpha w' \sim^{\overline{\alpha}} w'' \sim^\alpha w, w'\!:\!0, w''\!:\!1, w \sim^{\overline{\alpha}} w_1, w\!:\!1_\alpha, \Gamma \Rightarrow \Delta}{w \sim^{\overline{\alpha}} w_1, w\!:\!1_\alpha, \Gamma \Rightarrow \Delta}\;\text{BND}_\alpha$$

$$\frac{w \sim^\alpha w' \sim^{\overline{\alpha}} w'', w'' \succeq^\alpha w, w'\!:\!0, w''\!:\!1, w\!:\!0_\alpha, \Gamma \Rightarrow \Delta}{w\!:\!0_\alpha, \Gamma \Rightarrow \Delta}\;\text{NBD}_\alpha$$

- w', w'' are fresh in BND$_\alpha$ and NBD$_\alpha$.

Fig. 3. G3STIT$_2^d$

4.1 Structural properties

Given that all the rules extending the system follow the geometric pattern, the structural rules of the new system are straightforwardly established, using the familiar sequence below, expanding upon the proofs of [16,20].

- Derivability of sequents of the form $w\!:\!A, \Gamma \Rightarrow \Delta, w\!:\!A$ for any A.
- Height-preserving substitution on labels and agents.
- Height-preserving admissibility of weakening.

[11] Since i and j are stipulated to be distinct agents, the Difference rule from [16] is no longer required.

[12] Note that the system in [16] did not contain primitive $\Diamond^\alpha$. However, in parallel with $\mathcal{S}/\mathcal{P}$ and anticipating that this formulation will streamline future axiomatizations, we have decided to include it here.

- Height-preserving invertibility of all the rules.
- Height-preserving admissibility of contraction.
- Admissibility of cut.

We will briefly go over each in turn, noting first that the weight of the formula is defined in the standard way as

Definition 4.1 (Weight of a formula, $\mathtt{w}$) *The* weight of a labelled formula *$w{:}A$ is given by the weight of A, $\mathtt{w}(A)$, and is defined recursively as follows:*

- $\mathtt{w}(P) = \mathtt{w}(\bot) = \mathtt{w}(1_\epsilon) = \mathtt{w}(0_\epsilon) = 1,$
- $\mathtt{w}(A \supset B) = \mathtt{w}(A) + \mathtt{w}(B) + 1$,
- $\mathtt{w}(\Box^\alpha A) = \mathtt{w}(\Diamond^\alpha A) = \mathtt{w}(\mathcal{S}A) = \mathtt{w}(\mathcal{P}A) = \mathtt{w}(\neg A) = \mathtt{w}(A) + 1.$

Lemma 4.2 (Initial sequent generalization) *The sequents of the form $w{:}A, \Gamma \Rightarrow \Delta, w{:}A$ are derivable in G3STIT$^d_2$ for any formula A of $\mathfrak{L}$.*

Proof. Routine by induction on the weight of A. Since G3stit in [16] did not contain $\Diamond^\alpha$, we illustrate the case for it:

$$\dfrac{\dfrac{\dfrac{\text{i.h.}}{w \sim^\alpha w', w'{:}A, \Gamma \Rightarrow \Delta, w{:}\Diamond^\alpha A, w'{:}A}}{w \sim^\alpha w', w'{:}A, \Gamma \Rightarrow \Delta, w{:}\Diamond^\alpha A}\,\mathrm{R}\Diamond^\alpha}{w{:}\Diamond^\alpha A, \Gamma \Rightarrow \Delta, w{:}\Diamond^\alpha A}\,\mathrm{L}\Diamond^\alpha$$

The remaining new vocabulary (i.e. the constants) simply falls under the basic case of an initial sequent. □

Lemma 4.3 (Substitution) *If $\vdash_n \Gamma \Rightarrow \Delta$ is derivable in G3STIT$^d_2$, so are $\vdash_n \Gamma' \Rightarrow \Delta'$ and $\vdash_n \Gamma'' \Rightarrow \Delta''$, obtained from $\Gamma \Rightarrow \Delta$ by uniform substitution of labels and agent indices, respectively.*

Proof. By induction on the height of the derivation, using the inductive hypothesis twice when a clash of eigenvariables needs to be avoided. □

Lemma 4.4 (Weakening) *Weakening is hp-admissible: if $\vdash_n \Gamma \Rightarrow \Delta$ then $\vdash_n A, \Gamma \Rightarrow \Delta$ and $\vdash_n \Gamma \Rightarrow \Delta, B$, where A is a labelled formula or a relational atom, and B a labelled formula.*

Proof. Routine by induction on the height of the derivation, using Lemma 4.3 to avoid eigenvariable clashes. □

Lemma 4.5 (Invertibility) *All the rules of G3STIT$^d_2$ are hp-invertible: if the conclusion of the rule is derivable, so are its premises.*

Proof. By induction on the height of a derivation. Note that the proofs for the rules of G3STIT$_2$ are identical to those in [16] or straightforward for the new ones, while for all the deontic rules this is simply an application of Lemma 4.4 (hp-weakening). □

Lemma 4.6 (Contraction) *Contraction is hp-admissible in G3STIT$^d_2$: if $\vdash_n A, A, \Gamma \Rightarrow \Delta$ then $\vdash_n A, \Gamma \Rightarrow \Delta$ and if $\vdash_n \Gamma \Rightarrow \Delta, A, A$ then $\vdash_n \Gamma \Rightarrow \Delta, A$.*

Proof. By simultaneous induction on the height of the derivation. This is routine via the inductive hypothesis and possibly Lemma 4.5 applied to the premises of the last rule used. Note that the closure condition is met, namely whenever two principal formulas of a rule could be one and the same, the contracted version also appears as a rule (this is possible with $\mathit{Etrans}_{\sim^\alpha}$ and $\mathit{Repl}_{=}$, with the contracted versions being instances of $\mathit{Refl}_{\sim^\alpha}$ and $\mathit{Refl}_{=}$, respectively). □

Theorem 4.7 (Cut) *The rule of Cut is admissible in G3STIT$^{\mathrm{d}}_2$:*

$$\dfrac{\Gamma_1 \Rightarrow \Delta_1, C \qquad C, \Gamma_2 \Rightarrow \Delta_2}{\Gamma_1, \Gamma_2 \Rightarrow \Delta_1, \Delta_2}\ \mathit{Cut}$$

Proof. By induction on the weight of the cut formula with a secondary induction on the height of the cut (sum of the heights of its premises). Most of the proof is routine, and we will just illustrate the example of the cut formula of the form $\Diamond^\alpha A$ and principal in both premises. The instance of the Cut rule then has the form:

$$\dfrac{\dfrac{w \sim^\alpha w', \Gamma_1 \Rightarrow \Delta_1, w : \Diamond^\alpha A, w' : A}{w \sim^\alpha w', \Gamma_1 \Rightarrow \Delta_1, w : \Diamond^\alpha A}\ \mathrm{R}\Diamond^i \qquad \dfrac{w \sim^\alpha w', w' : A, \Gamma_2 \Rightarrow \Delta_2}{w : \Diamond^\alpha A, \Gamma_2 \Rightarrow \Delta_2}\ \mathrm{L}\Diamond^i}{w \sim^\alpha w', \Gamma_1, \Gamma_2 \Rightarrow \Delta_1, \Delta_2}\ \mathrm{Cut}$$

This is transformed into:

$$\dfrac{\dfrac{\dfrac{w \sim^\alpha w', \Gamma_1 \Rightarrow \Delta_1, w : \Diamond^\alpha A, w' : A \qquad w : \Diamond^\alpha A, \Gamma_2 \Rightarrow \Delta_2}{w \sim^\alpha w', \Gamma_1, \Gamma_2 \Rightarrow \Delta_1, \Delta_2, w' : A}\ \mathrm{Cut}_2 \qquad w \sim^\alpha w', w' : A, \Gamma_2 \Rightarrow \Delta_2}{w \sim^\alpha w', w \sim^\alpha w', \Gamma_1, \Gamma_2, \Gamma_2 \Rightarrow \Delta_1, \Delta_2, \Delta_2}\ \mathrm{Cut}_1}{w \sim^\alpha w', \Gamma_1, \Gamma_2 \Rightarrow \Delta_1, \Delta_2}\ \text{Lemma 4.6}$$

where Cut_1 is of lower weight, and Cut_2 of lower height, and therefore eliminable by primary and secondary inductive hypotheses, respectively.

Note that all the deontic rules follow the geometric rule schema, and therefore do not hinder the admissibility of cut. Specifically, since no relational formula or labelled atom is ever principal in a right rule, we can by reduction of height on the left obtain a premise of cut which is initial, in which case the cut is routinely eliminated. □

5 Soundness and completeness

In this section we establish that the proposed sequent calculus is sound and complete with respect to the Kripke semantics from Section 3. This also shows soundness and completeness w.r.t. the semantics from Section 2. Our overall proof methods follow [16], but their application to deontic admissibility constants is new to this paper.

5.1 Soundness of G3STIT$^{\mathrm{d}}_2$

In what follows M is assumed to be a relational model. We say that M makes $w : A$ true whenever $M, w \vDash A$, and that M makes a relational atom true iff the atom in question holds for M. We thus use truth to encompass claims that hold both *in* and *of* the model.

Theorem 5.1 (Soundness of G3STIT$^d_2$) *G3STIT$^d_2$ is sound: if a sequent $\Gamma \Rightarrow \Delta$ is derivable, then any relational model that makes all the formulas in Γ true also makes some formula in Δ true.*

Proof. By induction on the height of the derivation. Since for the rules of G3STIT$_2$ this is mostly a simplification of the proofs in [16], we only check for the deontic rules (L$\Diamond^\alpha$/R$\Diamond^\alpha$ are treated symmetrically to R$\Box^\alpha$ and L$\Box^\alpha$, respectively). Note that PB and Det straightforwardly correspond to, respectively, (D) and (Det) of relational models.

$01_\epsilon/\overline{01}_\epsilon$: Let $\Gamma \Rightarrow \Delta$ be derived by 01_ϵ:

$$\frac{w{:}1_\epsilon, \Gamma \Rightarrow \Delta \qquad w{:}0_\epsilon, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta}\; 01_\epsilon$$

Assume M makes all formulas in Γ true. Then by Definition 3.4, for every w, either $M, w \models 1_\epsilon$ or $M, w \models 0_\epsilon$. In either case some formula in Δ is true in M by the inductive hypothesis (IH). By the same Definition, either $M, w \not\models 1_\epsilon$ or $M, w \not\models 0_\epsilon$, so soundness trivially holds for any sequent derived by $\overline{01}_\epsilon$.

$\succeq^\alpha_L/\succeq^\alpha_R$: Let $\Gamma \Rightarrow \Delta$ be derived by $\succeq^\alpha_L$:

$$\frac{w \sim^\alpha w', w \succeq^\alpha w'', w' \succeq^\alpha w'', \Gamma \Rightarrow \Delta}{w \sim^\alpha w', w \succeq^\alpha w'', \Gamma \Rightarrow \Delta}\; \succeq^\alpha_L$$

Assume M makes all formulas in $w \sim^\alpha w', w \succeq^\alpha w'', \Gamma$ true. From the second atom it follows that $|w|_\alpha \succeq^\alpha |w''|_\alpha$. But since $w \sim^\alpha w'$, it follows that $|w|_\alpha = |w'|_\alpha$. So, $|w'|_\alpha \succeq^\alpha |w''|_\alpha$, and thus $w' \succeq^\alpha w''$. Then all of $w \sim^\alpha w', w \succeq^\alpha w'', w' \succeq^\alpha w'', \Gamma$ are true in M and by the IH so is some formula in Δ. The proof for $\succeq^\alpha_R$ is very similar and safely left to the reader.

BA$_\alpha$: Let $\Gamma \Rightarrow \Delta$ be derived by BA$_\alpha$:

$$\frac{w'{:}1_\alpha, w{:}1_\alpha, w \sim^\alpha w', \Gamma \Rightarrow \Delta}{w{:}1_\alpha, w \sim^\alpha w', \Gamma \Rightarrow \Delta}\; \mathrm{BA}_\alpha$$

Assume M makes all the formulas in $w{:}1_\alpha, w \sim^\alpha w', \Gamma$ true. Let $Y = |w|_\alpha = |w'|_\alpha \in \mathsf{A}_\alpha(M)$. Then, since $X \cap V(1_\alpha) \neq \emptyset$ and since M satisfies (BA$_\alpha$) (cf. Lemma 3.5), it follows that $X \subseteq V(1_\alpha)$. So, $w' \in V(1_\alpha)$ and hence $M, w' \models 1_\alpha$. In sum, all of $w'{:}1_\alpha, w{:}1_\alpha, w \sim^\alpha w', \Gamma$ are true in M, and by the IH so is some formula in Δ.

$\succeq^\alpha$: Let $\Gamma \Rightarrow \Delta$ be derived by $\succeq^\alpha$:

$$\frac{w \succeq^\alpha w', w \sim^{\overline{\alpha}} w', w'{:}1, w{:}1, \Gamma \Rightarrow \Delta}{w \succeq^\alpha w', w \sim^{\overline{\alpha}} w', w'{:}1, \Gamma \Rightarrow \Delta}\; \succeq^\alpha$$

Assume M makes all the formulas in $w \succeq^\alpha w', w \sim^{\overline{\alpha}} w', w'{:}1, \Gamma$ true. Let $Y = |w|_{\overline{\alpha}} = |w'|_{\overline{\alpha}} \in \mathsf{A}_{\overline{\alpha}}(M)$. From $w \succeq^\alpha w'$ it follows that $|w|_\alpha \succeq^\alpha |w'|_\alpha$, and therefore for every $Z \in \mathsf{A}_{\overline{\alpha}}(M) : d(|w|_\alpha \cap Z) \geq d(|w'|_\alpha \cap Z)$. Therefore, also $d(|w|_\alpha \cap Y) \geq d(|w'|_\alpha \cap Y)$. By Determinism, $d(w) \geq d(w')$. Since $w' : 1$ it follows that $w' \notin V(0)$. Thus, $w \in V(1)$, and so $w : 1$ holds. In sum, all of $w \succeq^\alpha w', w \sim^{\overline{\alpha}} w', w'{:}1, w{:}1, \Gamma$ are true, and by the IH so is some formula in

Δ.

BND$_\alpha$: Let $\Gamma \Rightarrow \Delta$ be derived by BND$_\alpha$:

$$\dfrac{w \succeq^\alpha w_1, w \sim^{\overline{\alpha}} w_1, w : 1_\alpha, \Gamma \Rightarrow \Delta \qquad w_1 \sim^\alpha w' \sim^{\overline{\alpha}} w'' \sim^\alpha w, w' : 0, w'' : 1, w \sim^{\overline{\alpha}} w_1, w : 1_\alpha, \Gamma \Rightarrow \Delta}{w \sim^{\overline{\alpha}} w_1, w : 1_\alpha, \Gamma \Rightarrow \Delta}\ \mathrm{BND}_\alpha$$

Assume M makes all the formulas in $w \sim^{\overline{\alpha}} w_1, w : 1_\alpha, \Gamma$ true. Case 1: $w_1 \succeq^\alpha w$. Given that $|w_1|_\alpha \succeq^\alpha |w|_\alpha$ and by (BA$_\alpha$), $|w|_\alpha \subseteq V(1_\alpha)$. Since M satisfies (BND$_\alpha$) (cf. Lemma 3.5), also $|w|_\alpha \succeq^\alpha |w_1|_\alpha$ and therefore $w \succeq^\alpha w_1$. In sum, all of $w \succeq^\alpha w_1, w \sim^{\overline{\alpha}} w_1, w : 1_\alpha, \Gamma$ are true and therefore by the IH so is some formula in Δ.
Case 2: $w_1 \not\succeq^\alpha w$. Then there is some $Y \in \mathsf{A}_{\overline{\alpha}}(M) : d(|w_1|_\alpha \cap Y) \not\geq d(|w|_\alpha \cap Y)$, i.e. such that $d(|w_1|_\alpha \cap Y) = 0$ and $d(|w|_\alpha \cap Y) = 1$. Let simply $w' \in |w_1|_\alpha \cap Y$ and $w'' \in |w|_\alpha \cap Y$, otherwise use Lemma 4.3 to (hp-)derive the sequent with those labels. Then, given Determinism, $d(w') = 0$ and $d(w'') = 1$, and therefore $w' : 0$ and $w'' : 1$ are both true. Moreover, since $w', w'' \in Y$, it holds that $w' \sim^{\overline{\alpha}} w''$. Since $w' \in |w_1|_\alpha$ it holds that $w_1 \sim^\alpha w'$ and since $w'' \in |w|_\alpha$ it holds that $w \sim^\alpha w''$. In sum, all of $w_1 \sim^\alpha w' \sim^{\overline{\alpha}} w'' \sim^\alpha w, w' : 0, w'' : 1, w \sim^{\overline{\alpha}} w_1, w : 1_\alpha, \Gamma$ are true, and by the IH so is some formula in Δ.

NBD$_\alpha$: Let $\Gamma \Rightarrow \Delta$ be derived by NBD$_\alpha$:

$$\dfrac{w \sim^\alpha w' \sim^{\overline{\alpha}} w'', w'' \succeq^\alpha w, w' : 0, w'' : 1, w : 0_\alpha, \Gamma \Rightarrow \Delta}{w : 0_\alpha, \Gamma \Rightarrow \Delta}\ \mathrm{NBD}_\alpha$$

Assume M makes all the formulas in $w : 0_\alpha, \Gamma$ true. Since M satisfies (BA$_\alpha$), it follows that $|w|_\alpha \subseteq V(0_\alpha)$. Since M satisfies (NBD$_\alpha$), there is a $Y \in \mathsf{A}_\alpha(M)$ s.t. $Y \succeq^\alpha |w|_\alpha$ and $|w|_\alpha \not\succeq^\alpha Y$. It follows from the latter that for some $Z \in \mathsf{A}_{\overline{\alpha}}(M) :\ d(|w|_\alpha \cap Z) \not\geq d(Y \cap Z)$, and so $w_0 \in |w|_\alpha \cap Z : w_0 \in V(0)$ and $w_1 \in Y \cap Z : w_1 \in V(1)$. Let $w' = w_0$ and $w'' = w_1$, otherwise use Lemma 4.3. So, $w' : 0$ and $w'' : 1$ both hold. Moreover, since $w_1 \in Y$, it follows that $w'' \succeq^\alpha w$. Since $w_0, w_1 \in Z \in \mathsf{A}_{\overline{\alpha}}(M)$, it follows that $w' \sim^{\overline{\alpha}} w''$. Finally, since $w_0 \in |w|_\alpha$, it follows that $w \sim^\alpha w'$. In sum, all of $w \sim^\alpha w' \sim^{\overline{\alpha}} w'', w'' \succeq^\alpha w, w' : 0, w'' : 1, w : 0_\alpha, \Gamma$ are true, and by the IH so is some formula in Δ. □

5.2 Completeness of G3STIT$_2^{\mathbf{d}}$

Completeness of G3STIT$_2^d$ is demonstrated by a countermodel construction via a failed proof search [19,14]. We begin by defining a reduction tree, which corresponds to a bottom-up proof search [23].

Definition 5.2 (Reduction tree) *A* reduction tree *for the sequent $\Gamma \Rightarrow \Delta$ is a tree built bottom-up in steps, each consisting of stages for each of the rules.*

A sequent that does not contain the same labelled atom in both the succedent and the consequent, or a bottom in the antecedent, or both $h : 1_\epsilon$ and $h : 0_\epsilon$ (for some label h) in the antecedent, is called active.

At each stage we apply, bottom-up, the rule of that stage to each leaf of the tree which is an active sequent $\Gamma_n \Rightarrow \Delta_n$. An application of a rule to a formula is called the reduction *of the latter. We don't reduce the formulas if the active formulas are already in the sequent. Once the rule of the stage can no longer* *be applied, we move to the next stage if there are still active sequents.*

The order of the stages in a step is the order of presentation of the rules in Figures 2 and 3, left to right and consecutively. For each rule with a freshness condition, we take from the denumerable list of labels the first label(s) not yet used in the tree. We apply $Refl_=$, $Refl_\alpha$ and 01_ϵ for any label occurring in $\Gamma_i \Rightarrow \Delta_i$ just once per branch, and we apply PB only once per branch.

If there are still active sequents once every stage of a step has been completed, a new step begins with the stage for propositional rules.

If we reach a point where no leaf is an active sequent, we have a derivation of $\Gamma \Rightarrow \Delta$. Namely, every leaf is either an initial sequent or a conclusion of $\overline{01}_\epsilon$, each sequent is generated by an application of a rule, and the endsequent is $\Gamma \Rightarrow \Delta$. Otherwise, we have a branch where no more steps can be applied, but all sequents are active. We use that branch to generate a countermodel.

Definition 5.3 (Refutation model M^c) *Let $\Gamma_0 \Rightarrow \Delta_0, \Gamma_1 \Rightarrow \Delta_1, \ldots$ be a (finite or infinite) branch of a reduction tree for $\Gamma \Rightarrow \Delta$ (so $\Gamma_0 \Rightarrow \Delta_0$ is just $\Gamma \Rightarrow \Delta$). Let $\Gamma^* = \bigcup \Gamma_{n\geq 0}$ and $\Delta^* = \bigcup \Delta_{n \geq 0}$. Let $M^c = \langle W, \sim^i, \sim^j, d, V\rangle$ and $I : \mathfrak{H} \to W$ be such that (i) $I(w) \in W$ iff w occurs in $\Gamma^* \cup \Delta^*$; (ii) $I(w) = I(w')$ iff $w = w'$ occurs in Γ^*; (iii) $I(w) \sim^\alpha I(w')$ iff $w \sim^\alpha w'$ occurs in Γ^*; and (iv) for every labelled atom $w : \varphi$: if $\varphi \in \mathfrak{P} \cup \{1_i, 1_j, 0_i, 0_j\}$, then $I(w) \in V(\varphi)$ iff $w : \varphi \in \Gamma^*$, if $\varphi = 1$, then $d(I(w)) = 1$ iff $w : \varphi \in \Gamma^*$, and if $\varphi = 0$, then $d(I(w)) = 1$ iff $w : \varphi \in \Delta^*$.*

Lemma 5.4 *M^c is a relational quasi-model.*

Proof. First, we show that $M^c, w \vDash 0_\epsilon$ iff $M^c, w \nvDash 1_\epsilon$. L-R holds because all sequents in the branch are active. R-L holds because by Definition 5.2 rule 01_ϵ has been applied to the branch in question.

Next, we show that in M^c both $\sim^\alpha$ are equivalence relations, because the rules $Refl_{\sim^\alpha}$ and $Etrans_{\sim^\alpha}$ have been applied to any (appropriate combination of) labels occurring in the branch. E.g. for every label w, $w \sim^\alpha w$ appears in the branch, and therefore $I(w) \in |I(w)|_\alpha$. Likewise for (IOA), given the rule Ind, and (Det), given rule Det. Finally, since in every branch the rule PB has been applied, for some $I(w) \in W$, $d(I(w)) = 1$, so W is non-empty and (D) holds. □

Lemma 5.5 *If $w \succeq^\alpha w'$ occurs in Γ^*, then $|I(w)|_\alpha \succeq^\alpha |I(w')|_\alpha$ in M^c.*

Proof. Let $w \succeq^\alpha w'$ occur in Γ^*. It then holds for every $w_1 \sim^\alpha w$ (and thus $I(w_1) \in |I(w)|_\alpha$) and every $w_2 \sim^\alpha w'$ (and thus $I(w_2) \in |I(w')|_\alpha$), since rules $\succeq^\alpha_L$ and $\succeq^\alpha_R$ have been respectively applied, that also $w_1 \succeq^\alpha w_2$ occurs in Γ^*.

Now let w_1 and w_2 be such a pair, and let $w_1 \sim^{\overline{\alpha}} w_2$ also occur in Γ^* (and therefore $I(w_1), I(w_2) \in Y \in \mathsf{A}_{\overline{\alpha}}(M^c)$). Since the rule $\succeq^\alpha$ has been applied it holds that if $d(I(w_2)) = 1$ then $d(I(w_1)) = 1$, and therefore $d(I(w_1)) \geq d(I(w_2))$. Since by Lemma 5.4 Determinism holds, it holds that for every $Z \in \mathsf{A}_{\overline{\alpha}}(M^c)$, $d(|I(w)|_\alpha \cap Z) \geq d(|I(w')|_\alpha \cap Z)$, i.e. $|I(w)|_\alpha \succeq^\alpha |I(w')|_\alpha$. □

Lemma 5.6 *M^c is a relational model.*

Proof. Given Lemmas 3.5 and 5.4, we just need to show that M^c satisfies (BA_α), (BND_α) and (NBD_α).

Ad (BA_α): Suppose that $X \in \mathsf{A}_\alpha(M^c)$ and $X \not\subseteq V(0_\alpha)$. Then for some $I(w) \in X$: $I(w) \in V(1_\alpha)$ and by Lemma 5.4, $X = |I(w)|_\alpha$. Let $I(w') \in |I(w)|_\alpha$. By Definition 5.3, $w : 1_\alpha$ and $w \sim^\alpha w'$ occur in Γ^*. Since the rule BA_α has been applied, $w' : 1_\alpha$ likewise occurs in Γ^*, and therefore $I(w') \in V(1_\alpha)$. Therefore, $|I(w)|_\alpha \subseteq V(1_\alpha)$ i.e. $X \subseteq V(1_\alpha)$. Hence either $X \subseteq V(0_\alpha)$ or $X \subseteq V(1_\alpha)$.

Ad (BND_α): Suppose that $|I(w)|_\alpha \subseteq V(1_\alpha)$. By Definition 5.3, $w : 1_\alpha$ occurs in Γ^*. Since the rule Ind has been applied and therefore relational atoms $w \sim^\alpha w_2 \sim^{\overline{\alpha}} w_1$ occur in Γ, and the rule BA_α has been applied and therefore the labelled formula $w_2 : 1_\alpha$ likewise occurs in Γ^*, the rule BND_α has also been applied. So, one of two things holds for every $|I(w_1)|_\alpha$.

First, that $w_2 \succeq^\alpha w_1$ occurs in Γ^*, and thus $|I(w_2)|_\alpha \succeq^\alpha |I(w_1)|_\alpha$ holds by Lemma 5.5. Otherwise there are some $I(w'), I(w'') \in Y \in \mathsf{A}_{\overline{\alpha}}(M^c)$ (from $w' \sim^{\overline{\alpha}} w''$) s.t. $I(w') \in |I(w_1)|_\alpha, V(0)$ (from $w_1 \sim^\alpha w'$ and $w' : 0$, respectively) and $I(w'') \in |I(w_2)|_\alpha, V(1)$ (similar). Therefore there is some $Y \in \mathsf{A}_{\overline{\alpha}}(M)$: $d(|I(w_1)|_\alpha \cap Y) \not\geq d(|I(w_2)|_\alpha \cap Y)$, i.e. such that $|I(w_1)|_\alpha \not\succeq^\alpha |I(w_2)|_\alpha$. Either way, since $|w|_\alpha = |w_2|_\alpha$ (Lemma 5.4, Definition 5.3 from $w \sim^\alpha w_2$), for every $|I(w)|_\alpha \subseteq V(1_\alpha)$, if $|I(w_1)|_\alpha \succeq^\alpha |I(w)|_\alpha$, then $|I(w)|_\alpha \succeq^\alpha |I(w_1)|_\alpha$.

Ad (NBD_α): Suppose that $|I(w)|_\alpha \subseteq V(0_\alpha)$. By Definition 5.3, $w : 0_\alpha$ occurs in Γ^*. Since NBD_α has been applied, for some label w'' the relational atom $w'' \succeq^\alpha w$ likewise occurs in Γ^*, and therefore by Lemma 5.5 it holds that $|I(w'')|_\alpha \succeq^\alpha |I(w)|_\alpha$.

Moreover, there is some $I(w'') \in |I(w'')|_\alpha$ s.t. $d(I(w'')) = 1$ (from $w'' : 1$, by Definition 5.3) and some $I(w') \in |I(w)|_\alpha$ (since $w \sim^\alpha w'$ occurs in Γ^*) s.t. $d(I(w')) = 0$ (from $w' : 0$). Therefore $d(I(w')) \not\geq d(I(w''))$ for some $I(w'), I(w'') \in Z \in \mathsf{A}_{\overline{\alpha}}(M^c)$ (since $w' \sim^{\overline{\alpha}} w''$ occurs in Γ^*), and hence $|I(w)|_\alpha \not\succeq^\alpha |I(w'')|_\alpha$. □

Lemma 5.7 (Refutation in M^c) *M^c makes any labelled formula A in Γ^* true and any labelled formula B in Δ^* false.*

Proof. By induction on the weight of the formula. The base case is covered by Definition 5.3 and Lemma 5.6. Most of the induction steps are familiar from [16], so we just cover the new case as an illustration.

So, let A be $w : \Diamond^\alpha C$. By Definition 5.2, for some w' s.t. $w \sim^\alpha w'$ (and thus $I(w') \in |I(w)|_\alpha$), $w' : C$ is in Γ^*, and by the inductive hypothesis $M^c, I(w') \vDash C$. Therefore $M^c, I(w) \vDash \Diamond^\alpha C$.

Let B be $w : \Diamond^\alpha C$. By Definition 5.2, for every w' s.t. $w \sim^\alpha w'$ (and thus $I(w') \in |I(w)|_\alpha$), $w' : C$ is in Δ^*, and by the inductive hypothesis $M^c, I(w') \nvDash C$. Therefore $M^c, I(w) \nvDash \Diamond^\alpha C$. □

Theorem 5.8 (Completeness of G3STIT$^d_2$) *G3STIT$^d_2$ is complete: if every relational model that makes all the formulas in Γ true also makes some formula in Δ true, then the sequent $\Gamma \Rightarrow \Delta$ is derivable.*

Proof. Suppose $\Gamma \Rightarrow \Delta$ is not derivable in G3STIT$^d_2$. Using Definition 5.3, we build a countermodel for this sequent. By Lemma 5.6 this model is a relational model. By Lemma 5.7 it validates all formulas in Γ (since they are all in Γ^*) and invalidates all formulas in Δ (all in Δ^*). So, if $\Gamma \Rightarrow \Delta$ is not derivable then there is a relational countermodel. □

6 Further work

Our proof theory suggests a more general recipe for proof theoretic characterizations of weak dominance in richer (cooperative or non-cooperative) games and for richer languages. More precisely, although they require some modelling decisions and additional technical machinery, each of the following extensions and generalizations are arguably within reach.

Smoothness of simple dominance. In previous work on deontic game models [26,25,5] the sets A_α are assumed to be finite. While finiteness is hard if not impossible to characterize[13] unless one imposes a fixed bound k on $|W|$, it is relatively easy to characterize the condition of *smoothness*: whenever some action X of α is not admissible, then there is some action X' of α that is admissible, with $X' \succeq^\alpha X$. Given this condition, it is easy to see that the strengthening of NBD_α where $w'' : 1_\alpha$ is added to the premise is sound, and gives us a complete calculus which retains all the structural properties.

The n-agent case. In order to generalize the sequent calculi to the case of $n \in \mathbb{N}$ distinct agents, note that in this setting, the role that is now played by $\overline{\alpha}$ will be taken over by $N - \{\alpha\}$. That is, simple dominance over the actions of α is defined by quantifying over all combinations of actions by all the agents in $N - \{\alpha\}$. Consequently, in the rules $\succeq^\alpha$, BND_α, and NBD_α, the expression $w \sim^{\overline{\alpha}} w'$ should be read as: for every $\beta \in N - \{\alpha\}$, $w \sim^\beta w'$. Similarly, the rule Det that expresses the determinism of the model should be rewritten so that worlds are identical if and only if they are in the same equivalence class for all the agents. Finally, Ind should be replaced with the like-named rule from [16], with the difference of agents reintroduced.

Agent-relative standards of admissibility. In the preceding, we presupposed that there is a unique, common normative evaluation of the possible worlds. In game-theoretic terms, we focused on cooperative games. Instead, one may also consider agent-dependent acceptability constants d_α (for "as far as α is concerned, this is an acceptable world") and their negation v_α. An adequate proof system for such a richer logic would be obtained by replacing 1 and 0 with those two respective constants. Note that this allows us to model both notions of egoistic admissibility – i.e. actions of α are admissible for α if and only if they are not weakly dominated with respect to α's own standards – and altruistic admissibility – i.e. actions of α are admissible for α iff they are not weakly dominated with respect to $\overline{\alpha}$'s standards. More generally, any boolean combination of (agent-relative or agent-independent) deontic constants may be taken as the evaluative standard for admissibility.

Non-binary deontic evaluation. Current work on deontic game models presupposes a simple, binary classification of action profiles into acceptable and

[13] See also footnote 6.

unacceptable ones. A natural extension would allow for some finite number k of distinct deontic values that are used to evaluate the outcomes of combined actions. To obtain a proof theoretic characterization of admissibility in this setting, one may draw on [4], where propositional constants u_n are used to express that the given world has a utility (deontic value) of at least n. A rule such as our $\succeq^\alpha$ would then have to be rewritten, replacing 1 with u_n for every $n \in \{1, \ldots, k\}$. While we think that such a richer semantics can be adequately characterized by an extension of our sequent calculi, we leave the full exploration of this and the other mentioned variants for future work.

References

[1] Anderson, A. R., *A reduction of deontic logic to alethic modal logic*, Mind **LXVII** (1958), pp. 100–103.
URL `https://doi.org/10.1093/mind/LXVII.265.100`

[2] Badura, C. and H. Wansing, *STIT-logic for imagination episodes with voluntary input*, The Review of Symbolic Logic (2021), pp. 1–49.

[3] Belnap, N. D., M. Perloff and M. Xu, "Facing the Future: Agents and Choice in Our Indeterminist World," Oxford University Press, 2001.

[4] De Coninck, Thijs and Van De Putte, Frederik, *The original position : a logical analysis*, in: Liu, Fenrong and Marra, Alessandra and Portner, Paul and Van De Putte, Frederik, editor, *Deontic logic and normative systems : 15th international conference, DEON 2020/2021* (2021), pp. 133–150.

[5] Duijf, H., A. Tamminga and F. Van De Putte, *An impossibility result on methodological individualism*, Philosophical Studies **178** (2021), pp. 4165–4185.

[6] Duijf, H. and F. Van De Putte, *The problem of no hands: responsibility voids in collective decisions*, Social Choice and Welfare **58** (2022), p. 753–790.
URL `http://dx.doi.org/10.1007/s00355-021-01364-5`

[7] Gratzl, N. and E. Pavlović, *Is, ought, and cut*, Journal of Philosophical Logic (2023), pp. 1–21.

[8] Herrestad, H. and C. Krogh, *Obligations directed from bearers to counterparts*, in: *International Conference on Artificial Intelligence and Law*, 1995, pp. 210–218.

[9] Horty, J. F., "Agency and Deontic Logic," Oxford University Press, New York, 2001.

[10] Kanger, S., "New Foundations for Ethical Theory," Springer Netherlands, Dordrecht, 1971 [1957] pp. 36–58.
URL `https://doi.org/10.1007/978-94-010-3146-2_2`

[11] Lindahl, L., "Position and Change: A Study in Law and Logic," Dordrecht: D. Reidel., 1977.

[12] Lyon, T. and K. van Berkel, *Automating agential reasoning: Proof-calculi and syntactic decidability for STIT logics*, in: *International Conference on Principles and Practice of Multi-Agent Systems*, Springer, 2019, pp. 202–218.

[13] Murakami, Y., *Utilitarian deontic logic*, Advances in Modal Logic **5** (2005), pp. 211–230.

[14] Negri, S., *Proofs and countermodels in non-classical logics*, Logica Universalis **8** (2014), pp. 25–60.

[15] Negri, S. and R. Dyckhoff, *Geometrization of first-order logic*, Bulletin of Symbolic Logic **21** (2015), pp. 123–163.

[16] Negri, S. and E. Pavlović, *Proof-theoretic analysis of the logics of agency: The deliberative STIT*, Studia Logica **109** (2021), pp. 473–507.

[17] Negri, S. and E. Pavlović, *Alternative axiomatization for logics of agency in a G3 calculus*, Foundations of Science **28** (2023), pp. 205–224.

[18] Negri, S. and J. von Plato, *Cut elimination in the presence of axioms*, Bulletin of Symbolic Logic **4** (1998), p. 418–435.

[19] Negri, S. and J. Von Plato, “Structural proof theory,” Cambridge university press, Cambridge, 2001.

[20] Negri, S. and J. von Plato, “Proof Analysis: A Contribution to Hilbert’s Last Problem,” Cambridge University Press, 2011.

[21] Olkhovikov, G. and H. Wansing, *Simplified tableaux for STIT imagination logic*, Journal of Philosophical Logic **48** (2019), pp. 981–1001.

[22] Orlandelli, E., *Proof analysis in deontic logics*, , **12**, Springer, 2014, pp. 139–148.

[23] Pavlović, E. and N. Gratzl, *A more unified approach to free logics*, Journal of Philosophical Logic **50** (2021), pp. 117–148.

[24] Tamminga, A. and H. Duijf, *Collective obligations, group plans, and individual actions*, Economics and Philosophy **33** (2017), pp. 187–214.

[25] Tamminga, A., H. Duijf and F. Van De Putte, *Expressivity results for deontic logics of collective agency*, Synthese **198** (2021), pp. 8733–8753.

[26] Tamminga, A. and F. Hindriks, *The irreducibility of collective obligations*, Philosophical Studies **177** (2020), pp. 1085–1109.

[27] van Benthem, J. and D. Klein, *Logics for Analyzing Games*, in: E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, 2022, Winter 2022 edition .

[28] van Benthem, J. and E. Pacuit, “Connecting Logic of Choice and Change,” Outstanding Contributions to Logic **Nuel Belnap on indeterminism and free action**, Springer, 2014 pp. 291–314.

[29] Van Benthem, J., E. Pacuit and O. Roy, *Toward a theory of play: A logical perspective on games and interaction*, Games **2** (2011), pp. 52–86.
URL `https://www.mdpi.com/2073-4336/2/1/52`

[30] van Benthem, J., S. van Otterloo and O. Roy, *Preference logic, conditionals and solution concepts in games*, in: *Modality matters : twenty-five essays in honour of Krister Segerberg*, Uppsala Philosophical Studies **53**, Univ., Dept. of Philosophy, Uppsala, 2006 pp. 61–77.
URL `https://eref.uni-bayreuth.de/id/eprint/40308/`

[31] van Berkel, K. and T. Lyon, *Cut-free calculi and relational semantics for temporal STIT logics*, in: *European Conference on Logics in Artificial Intelligence*, Springer, 2019, pp. 803–819.

[32] van Berkel, K. and T. Lyon, *The varieties of ought-implies-can and deontic STIT logic*, in: F. Liu, A. Marra, P. Portner and F. V. D. Putte, editors, *Deontic Logic and Normative Systems: DEON 2021*, 2021 pp. 55–76.

[33] Van De Putte, F., A. Tamminga and H. Duijf, *Doing without nature*, in: A. Baltag, J. Seligman and T. Yamada, editors, *Proceedings of the Sixth International Workshop on Logic, Rationality, and Interaction* (2017), pp. 209–223.

A Dynamic Logic of the Right to Know

Xu Li[1] [2]

University of Luxembourg, Esch-sur-Alzette, Luxembourg

Réka Markovich[1] [2]

University of Luxembourg, Esch-sur-Alzette, Luxembourg

Abstract

Different meanings of the "right to know" can be distinguished based on the theory of normative positions. In this paper, we focus on one of them—the power to know. Intuitively, in a sender-receiver setting, the receiver's power to know whether φ is the case means that the sender is obliged to (truthfully) announce the answer if the receiver asks the question φ?. Therefore, we develop a logic called LRK for reasoning about the power to know, the obligatory announcements, and the dynamics of questions and public announcements. We explore some semantic results for LRK and study the expressive power of fragments of LRK. In particular, we show that no DEL-style reduction axiomatization exists for LRK.

Keywords: epistemic rights, public announcement logic, logic of questions

1 Introduction

The aim of this paper is to provide a logical framework to reason about the notion of the right to know and its interaction with other related notions. As a theoretical-conceptual background, we rely on the theory of normative positions [18] based on the theory of Hohfeld [9,15]. Using this analysis, we can differentiate between four different meanings of the right to know: the privilege to know, the claim-right to know, the power to know, and the immunity to know. The difference between the four meanings of the right to know can be illustrated by the following example [23]: suppose that an agent i has been tested for some disease by his doctor. It is usually mentioned in laws that i has the right to know the test result. However, it is unclear which type of right they mean by it. i has a privilege-right to know the result which means that i has no duty not to know them, while i has a claim-right to know the result meaning that his doctor has a duty to inform him about the result. Besides,

[1] {xu.li, reka.markovich}@uni.lu

[2] This work was supported by the Fonds National de la Recherche Luxembourg through the project Deontic Logic for Epistemic Rights (OPEN O20/14776480) and the project INTEGRAUTO (INTER/AUDACE/21/16695098).

i has a power-right to know the result meaning that his doctor has a duty to inform them if i requests it. Last, i has an immunity-right to know the results, which protects him from his doctor taking away or altering his claim-right to know the results.

In the logical literature, so far the claim-right to know received explicit attention, see, e.g., [16]. In this paper, we will investigate the right to know as a power. Power is characterized in [15] as a potential: the agent having the power is able to execute an action resulting in the counterparty's normative positions changing, e.g. a duty arising. Thus in this case: the patient's asking about the test results creates the doctor's duty to tell.

The notion of the power to know can also be found extensively in database theory under the name of "the right/permission to access" (see, e.g., [5]). The logical characterization of "the right/permission to access" is crucial for practical problems like maintaining security in databases. The point of the problem is that the database has to answer the users' queries while complying with certain security policies (e.g., privacy policies). Many factors affecting the solutions to the problem have been identified in the literature. E.g., the representation of the database, the expressiveness of the query language, the space of admissible responses, the initial knowledge of the user, etc [3]. However, the expressiveness of the language for specifying security policies has somehow been overlooked. To the best of our knowledge, all existing works on the topic consider the permitted, forbidden, and obligatory knowledge/belief of users as the only components of a security policy, see, e.g., [3,5,1]. However, as a counter-example, it is stated in the General Data Protection Regulation of Europe (GDPR) that "A data subject should have *the right of access* to personal data which have been collected concerning him or her". We illustrate by a toy example the role of "the right/permission to access" in solving the security problem for databases. Consider two security policies $\mathcal{P}_1$ and $\mathcal{P}_2$ where $\mathcal{P}_1$ consists of the following clauses (1) and (2), and $\mathcal{P}_2$ the clauses (1), (2), and (3).

(1) The user i is permitted to know whether p.
(2) The user i is forbidden to know whether q.
(3) The user i has the right to access the answer to whether p.

Suppose further that the user i knows that p is equivalent to q. If the user i asks the database whether p holds, there are two different situations. Under the security policy $\mathcal{P}_1$, the database has solutions to this query, e.g., the database may keep silent.[3] However, under the security policy $\mathcal{P}_2$, no solution exists because the database is forced to answer the query p?.

In this paper, we develop a Logic of the Right to Know, LRK, to reason about the notion of the power-type of right to know whether something is the case (we often will refer to it as a 'power to know' in what follows). As suggested

[3] A dispute may arise on whether the database may keep silent. This depends on our understanding of "permission to know": does the user's permission to know whether p imply that the database needs to answer the query p? raised by the user? But this is exactly a sign that we need different formalizations for these two different notions.

by the previous examples, the notion of the power to know is closely intertwined with other notions such as the obligation to inform, the public announcement, and the dynamics of questions. Therefore, LRK is devised such that these notions can also be expressed in the language.

The paper is structured as follows. In the next section, we introduce the language and semantics of LRK, as well as an example illustrating them. Section 3 and Section 4 are devoted to some semantic results of LRK and the expressive power of fragments of LRK, respectively. We discuss related literature in Section 5 and conclude with Section 6.

2 Language and Semantics

In this section, we introduce the language and semantics of LRK and illustrate them by an example. The scenarios that LRK is intended to characterize are communications between two agents where the information can only be transmitted from one agent (the sender/speaker, indicated by s) to the other (the receiver/addressee, indicated by r).The sender is further subject to some security policies such as privacy policies. These scenarios are common in our life, e.g., communications between a database and its users, conversations between a doctor and their patients, etc. We fix the role of the sender making only one of the agents able to make announcements. The restriction may seem to be unnatural. But, for simplicity, we will only consider the restricted scenarios. The following will serve as the running example of this paper:

Example 2.1 Two patients a and b have been tested for some diseases by the same doctor in the same time period. The policy consists of the following clauses: (1) The doctor is obliged to inform the patients about their test results; (2) Any patient is forbidden to know others' results; (3) The patients have the power to know whether the cheaper medicine is as good as the expensive one.

Suppose that, unfortunately, the test results for both a and b are positive and the cheaper medicine has the same efficacy as the expensive one. Let the sender be the doctor and the receiver the patient a. The doctor needs to decide which information should be informed.

Let us first introduce the formal language. Let PROP be a countable infinite set of propositional variables.

Definition 2.2 The language $\mathcal{L}$ is given by the following BNF grammar:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \to \varphi) \mid K_r\varphi \mid R_r\varphi \mid \mathbb{O}_s\varphi \mid \Box\varphi \mid [r : \varphi?]\varphi \mid [s : \varphi!]\varphi$$

where $p \in$ PROP. Other boolean connectives are defined as usual and $\langle s : \varphi! \rangle\psi$ is an abbreviation for $\neg[s : \varphi!]\neg\psi$.

The readings of the operators in $\mathcal{L}$ are as follows:

- $K_r\varphi$: The receiver knows φ.
- $R_r\varphi$: The receiver has the power-right to know the answer to the question φ? (or, equivalently, the receiver has the power-right to know whether φ).

- $\mathbb{O}_s\varphi$: The sender is obliged to (truthfully) announce φ.
- $\Box\varphi$: It is universally true that φ.
- $[r\!:\!\varphi?]\psi$: After the receiver asked the question $\varphi?$, it holds that ψ.
- $[s\!:\!\varphi!]\psi$: After the sender (truthfully) announced φ, it holds that ψ.

Next, we introduce the models for LRK. Our models are essentially a combination of the "neighbourhood epistemic models" introduced in [12] and the "epistemic issue models" in [21], see Section 5.

Definition 2.3 A model is a tuple $M = (W, \sim, \approx, N, V)$ where:

- W is a non-empty set of possible worlds or states;
- $\sim \subseteq W \times W$ is an equivalence relation on W;
- $\approx \subseteq \wp(W)$ is a partition of W;
- $N : W \to \wp(\wp(W))$ is such that $w \in U$ for all $w \in W$ and $U \in N(w)$;
- $V : \textsc{prop} \to \wp(W)$ is a valuation.

A pointed model is a pair M, w such that w is a state of M. For every state $w \in W$, we denote the set $\{v \in W \mid w \sim v\}$ as $\sim (w)$. We will also denote the unique subset in $\approx$ that contains w as $\approx (w)$.

In the above definition, $\sim$ is the familiar epistemic indistinguishability relation (of the receiver). The partition $\approx$ is intended to encode the set of questions to which the receiver has the power to know the answers. The set of questions may be stipulated by some normative systems such as privacy policies. The idea to represent a set of questions by a partition can be found in, e.g., [7], [4], and [21]. Finally, each subset in the neighbourhood $N(w)$ is an ideal epistemic state for the receiver at w. Thus the neighbourhood function N specifies which epistemic states (of the receiver) are ideal at every state w.

Let us illustrate the definition of the models by the running example:

Example 2.4 Let p_a, p_b, and g be the propositions that "The result for a is positive", "The result for b is positive", and "The cheaper medicine is as good as the expensive one", respectively. The case in Example 2.1 can be characterized by the model $M = (W, \sim, \approx, N, V)$ (as illustrated in Fig. 1) where:

- $W = \{000, 001, 010, 011, 100, 101, 110, 111\}$;
- $\sim = W \times W$;
- $\approx = \{\{000, 010, 100, 110\}, \{001, 011, 101, 111\}\}$;
- for all $xyz \in \{000, 001, 010, 011\}$, $N(xyz) = \{U \mid xyz \in U \,\&\, U \subseteq \{000, 001, 010, 011\} \,\&\, U \not\subseteq \{000, 001\} \,\&\, U \not\subseteq \{010, 011\}\}$,
  for all $xyz \in \{100, 101, 110, 111\}$, $N(xyz) = \{U \mid xyz \in U \,\&\, U \subseteq \{100, 101, 110, 111\} \,\&\, U \not\subseteq \{100, 101\} \,\&\, U \not\subseteq \{110, 111\}\}$;
- $V(p_a) = \{xyz \in W \mid x = 1\}$,
  $V(p_b) = \{xyz \in W \mid y = 1\}$,
  $V(g) = \{xyz \in W \mid z = 1\}$.

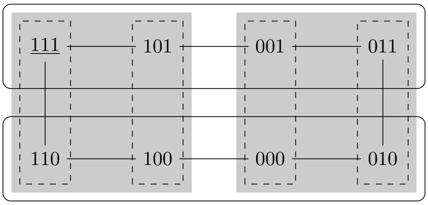


Fig. 1. The model $M = (W, \sim, \approx, N, V)$. States are 000, 001, ... and the actual state is 111. For each state xyz, $x = 1$ ($y = 1$, $z = 1$, respectively) iff $xyz \in V(p_a)$ ($xyz \in V(p_b)$, $xyz \in V(g)$, respectively). The indistinguishability relation $\sim$ is indicated by the straight line (with the reflexive and transitive arrows omitted). $\approx$ is the partition indicated by the rectangles with rounded corners (this captures the patient's power to know g). Finally, for every $xyz \in W$, $N(xyz)$ consists of the subsets such that: (1) it contains xyz itself; (2) it is contained in one of the shaded areas (this corresponds to the doctor's obligation to inform about p_a); (3) it is not contained in one of the dashed rectangles (the patient is prohibited to know p_b).

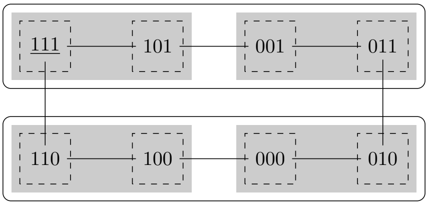


Fig. 2. The updated model $M_{g?}$. The same convention is adopted as in Fig. 1.

The next step is to provide the semantics for $\mathcal{L}$, especially for the formulas $R_r\varphi$ and $[r : \varphi?]\psi$. The semantics for $R_r\varphi$ is relatively straightforward: the receiver has the power to know the answer to φ? iff the question φ? is "settled" by the partition $\approx$, in the sense that all the cells in the partition $\approx$ are subsets of the truth set of φ or $\neg\varphi$. One may already notice that the truth of $R_r\varphi$ does not depend on the evaluating states. Hence, the notion of the power to know is a global notion in our models.[4] To express this fact, we use the universal modality $\Box$.

For the semantics for $[r : \varphi?]\psi$, as mentioned before, $[r : \varphi?]\psi$ is intended to express that "after the receiver asked the question φ?, it holds that ψ". If the receiver has no power to know the answer to φ?, then nothing will change after the action $[r : \varphi?]$. On the contrary, if the receiver indeed has the power, the sender is then forced to answer the question φ?. This is modeled in our

[4] It also makes sense to generalize our models to account for cases where, for example, the receiver is not aware of their rights to know. We leave that for future work.

framework in a way such that, in the updated model, all epistemic states that are not answers to $\varphi?$ become no longer ideal. Formally, the idea is captured by the definition of $M_{\varphi?}$ in the next definition.

Definition 2.5 Given a model $M = (W, \sim, \approx, N, V)$, for all $w \in W$ and $\varphi \in \mathcal{L}$, the satisfaction relation $M, w \models \varphi$ is inductively defined as follows:

$$\begin{array}{rcl}
M, w \models p & \text{iff} & w \in V(p) \\
M, w \models \neg\varphi & \text{iff} & M, w \not\models \varphi \\
M, w \models (\varphi \to \psi) & \text{iff} & M, w \not\models \varphi \text{ or } M, w \models \psi \\
M, w \models K_r\varphi & \text{iff} & \text{for all } v \in W,\ w \sim v \text{ implies } M, v \models \varphi \\
M, w \models R_r\varphi & \text{iff} & \text{for all } U \in \approx,\ U \subseteq [\![\varphi]\!]_M \text{ or } U \subseteq [\![\neg\varphi]\!]_M \\
M, w \models \mathbb{O}_s\varphi & \text{iff} & \text{for all } U \in N(w),\ U \subseteq \sim(w) \text{ implies } U \subseteq [\![\varphi]\!]_M \\
M, w \models \Box\varphi & \text{iff} & \text{for all } v \in W,\ M, v \models \varphi \\
M, w \models [r : \varphi?]\psi & \text{iff} & M_{\varphi?}, w \models \psi \\
M, w \models [s : \varphi!]\psi & \text{iff} & M, w \models \varphi \text{ implies } M_{\varphi!}, w \models \psi
\end{array}$$

where $[\![\varphi]\!]_M = \{x \in W \mid M, x \models \varphi\}$ is the truth set of φ in M and the models $M_{\varphi?}$ and $M_{\varphi!}$ are defined as follows:

$M_{\varphi?} = M$ if $M, w \not\models R_r\varphi$; otherwise $M_{\varphi?} = (W_{\varphi?}, \sim_{\varphi?}, \approx_{\varphi?}, N_{\varphi?}, V_{\varphi?})$ where:

- $W_{\varphi?} = W$, $\sim_{\varphi?} = \sim$, $\approx_{\varphi?} = \approx$, $V_{\varphi?} = V$,
- $N_{\varphi?}(x) = \{U \in N(x) \mid U \subseteq [\![\varphi]\!]_M \text{ or } U \subseteq [\![\neg\varphi]\!]_M\}$ for all $x \in W$.

$M_{\varphi!} = (W_{\varphi!}, \sim_{\varphi!}, \approx_{\varphi!}, N_{\varphi!}, V_{\varphi!})$ where:

- $W_{\varphi!} = W$, $\approx_{\varphi!} = \approx$, $N_{\varphi!} = N$, $V_{\varphi!} = V$,
- $\sim_{\varphi!} = \sim \cap \{(u, v) \in W \times W \mid M, u \models \varphi \text{ iff } M, v \models \varphi\}$.

The semantics for $K_r\varphi$ and $[s : \varphi!]\psi$ is standard, except that in the definition of $M_{\varphi!}$ we choose to delete the links between the φ-states and $\neg\varphi$-states, instead of removing all $\neg\varphi$-states from the model. Those $\neg\varphi$-states are reserved for further reference.[5] This kind of model updating can be found in, e.g., [20,21]. The intuition behind the semantics for $R_r\varphi$ and $[r : \varphi?]\psi$ has been explained. As for $\mathbb{O}_s\varphi$, it reflects the intuition that the sender is obliged to announce φ if φ holds in all ideal epistemic states (for the receiver) that are achievable by further announcements (i.e., φ is a piece of necessary information for restoring ideality). Let us illustrate the semantics by the running example:

Example 2.6 In the model M, we have, e.g., $M, 111 \models \mathbb{O}_s p_a$, $M, 111 \models R_r g$, and $M, 111 \not\models \mathbb{O}_s p_b$. After the receiver (the patient a) asked "Is the cheaper medicine as good as the expensive one?" ($g?$), the updated pointed model is $M_{g?}$ in Figure 2. We have, e.g., $M_{g?}, 111 \models \mathbb{O}_s p_a$, $M_{g?}, 111 \models R_r g$, and $M_{g?}, 111 \models \mathbb{O}_s g$.

A question may arise regarding the semantics for $\mathbb{O}_s\varphi$: why do we just consider ideal epistemic states achievable by further announcements instead of

[5] This deviates from the classical public announcement logic. But, in our case, deleting all $\neg\varphi$-states may change the truth of formulas like $R_r p$, which is unreasonable.

all? Technically, one may propose the following alternative semantic definition for $\mathbb{O}_s\varphi$:

$$M, w \models \mathbb{O}_s\varphi \quad \text{iff} \quad \text{for all } U \in N(w),\, U \subseteq [\![\varphi]\!]_M. \qquad (\dagger)$$

We will argue, however, that (†) does not work in certain cases within the context of database security, whereas our original semantics does. Consider the following example taken from [1]:

Example 2.7 Suppose the sender is communicating classified information to the receiver. Since the receiver is permitted to know some information p, the only constraint for the sender is that it is forbidden for the receiver to know p while not to know that it is classified (c). Furthermore, suppose that the receiver currently knows p, but she does not know c. The scenario can be represented by the following model $M = (W, \sim, \approx, N, V)$ such that:

- $W = \{w, u, v, x\}$;
- $\sim(w) = \sim(u) = \{w, u\}$, $\sim(v) = \{v\}$, and $\sim(x) = \{x\}$;
- $\approx = \{W\}$;
- $N(w) = \{Y \subseteq W \mid w \in Y \text{ and } Y \neq \{w, u\}\}$;
- $V(p) = \{w, u\}$ and $V(c) = \{w, v\}$.

Intuitively, we will agree that, in this case, it is obligatory for the sender to announce c. This is consistent with our semantics since $M, w \models \mathbb{O}_s c$. However, it would not be the case if we adopted (†).

When defining our models, the set of ideal epistemic states for a state w, $N(w)$, is not required to be non-empty. This is in contrast with standard deontic logic. Formally, we say a model $M = (W, \sim, \approx, N, V)$ is *standard* if for all $w \in W$, $N(w) \neq \varnothing$. The next proposition shows that the class of standard models gives the same logic as the class of all models.

Proposition 2.8 *For all pointed models M, w, there is a standard model M' and w' in M' such that $M, w \models \varphi$ iff $M', w' \models \varphi$ for all $\varphi \in \mathcal{L}$.*

Proof. Given a model $M = (W, \sim, \approx, N, V)$, we define the *double model* of M, $M' = (W', \sim', \approx', N', V')$, as follows:

- $W' = W \times \{1, 2\}$ (elements of W' will be denoted by $w_1, w_2, \dots$);
- $w_i \sim' v_j$ iff $w \sim v$ and $i = j$;
- $\approx' = \{U \times \{1\}, U \times \{2\} \mid U \in \approx\}$;
- for all $w_i \in W'$, $N'(w_i) = \{U \times \{i\} \mid U \in N(w)\} \cup \{\{w\} \times \{1, 2\}\}$;
- $V'(p) = V(p) \times \{1, 2\}$.

It is clear that $N'(w_i) \neq \varnothing$ for any $w_i \in W'$, hence M' is a standard model. Then the proposition follows from the following claim:

Claim. For all $w_i \in W'$, $M', w_i \models \varphi$ iff $M, w \models \varphi$ for all $\varphi \in \mathcal{L}$.
Proof of Claim. Induction on the structure of φ. Here we show only the inductive step for $\mathbb{O}_s\psi$: From left to right. Suppose $M', w_i \models \mathbb{O}_s\psi$. Let $U \in N(w)$ be such that $U \subseteq \sim(w)$. It suffices to show that $U \subseteq [\![\psi]\!]_M$. Note

that $U \times \{i\} \in N'(w_i)$ and $U \times \{i\} \subseteq\sim' (w_i)$ by the definition of N' and $\sim'$, respectively. Hence $U \times \{i\} \subseteq [\![\psi]\!]_{M'}$ by the assumption. It follows that $U \subseteq [\![\psi]\!]_M$ by the IH. From right to left. Suppose $M, w \models \mathbb{O}_s\psi$. Let $U' \in N'(w_i)$ be such that $U' \subseteq\sim' (w_i)$. It suffices to show that $U' \subseteq [\![\psi]\!]_{M'}$. Since $\{w\} \times \{1, 2\} \not\subseteq\sim' (w_i)$, there must be $U \in N(w)$ such that $U' = U \times \{i\}$. Note also that $U \subseteq\sim (w)$ since $U' \subseteq\sim' (w_i)$. Hence $U \subseteq [\![\psi]\!]_M$ by the assumption. Thus $U' \subseteq [\![\psi]\!]_{M'}$ by the IH.

For the case $[r : \psi?]\chi$ (and, similarly, $[s : \psi!]\chi$), note that $M'_{\psi?}$ is still the double model of $M_{\psi?}$ by the IH. Hence, by applying the IH, we have $M'_{\psi?}, w_i \models \chi$ iff $M_{\psi?}, w \models \chi$. □

3 Some Semantic Results

In this section, we list some (in)validities of LRK. The first group of validities is about the notion of the power to know. We omit the proofs because they are all straightforward.

Proposition 3.1 *The following hold for all formulas φ and ψ:*

(1) $\models R_r\top$.

(2) $\models R_r\varphi \wedge R_r\psi \to R_r(\varphi \wedge \psi)$.

(3) $\models R_r\varphi \to R_r\neg\varphi$.

(4) $\models \Box(\varphi \leftrightarrow \psi) \to (R\varphi \leftrightarrow R\psi)$.

From the above, we can see that the fragment of LRK on the notion of the power to know is nothing but the logical entailment relation between questions. That is to say, in LRK, if the answer to a question φ? can be derived from that of a set of questions that the receiver has the power to know the answers to, then the receiver also has the power to know the answer to φ?. [6]

Proposition 3.2 *The following hold for all formulas φ and ψ:*

(1) $\models K_r(\varphi \to \psi) \to (\mathbb{O}_s\varphi \to \mathbb{O}_s\psi)$.

(2) $\models \neg\mathbb{O}_s\bot \to (\mathbb{O}_s\varphi \to \varphi)$.

(3) $\models \mathbb{O}_s(\varphi \to \psi) \to (\mathbb{O}_s\varphi \to \mathbb{O}_s\psi)$.

(4) $\models \Box\varphi \to \mathbb{O}_s\varphi$.

(5) $\models K_r\varphi \to \mathbb{O}_s\varphi$.

The proofs are again omitted due to the same reason. A few remarks can be made on the above logical rules governing the behavior of the operator $\mathbb{O}_s$. The first says that if φ is more informative than ψ for the receiver and the sender is obliged to inform φ,[7] then the sender is also obliged to inform ψ.

[6] One may argue that the notion of the power to know characterized in LRK is a rather weak notion. For example, in most scenarios, the receiver has only the power to know what they are permitted to know (otherwise there would be conflict in the security policy). To model these scenarios, we can impose extra constraints on the models, e.g., for all $w \in W$, there is $U \in N(w)$ such that $\approx (w) \subseteq U$. Since the focus of this paper is on the formalization of the power to know, we leave this for future work.

[7] We follow [1, Definition 9] for the definition of "informativeness" of formulas.

The second says that the sender is not obliged to lie unless they are obliged to announce the contradiction (i.e., they face a deontic dilemma). Further, the validities (3) and (4) indicate that $\mathbb{O}_s\varphi$ is a normal modality. But we have a problem with interpreting the last validity. Literally, it states that the sender is obliged to announce whatever the receiver knows. This seems rather counterintuitive. To understand (5), note that, if φ is known by the receiver, the announcement of φ is actually less informative than any announcement for the receiver. Thus, from the informational point of view, the announcement of φ is "implied" by any announcement. In this sense, the announcement of φ is inevitable or necessary in our system since we assume that the sender can only make (truthful) announcements. So, the obligatory announcement of φ simply follows from the fact that the announcement of φ is necessary. [8]

The previous two propositions are about the properties of "the power to know" and "obligatory announcements" separately. However, it is natural to expect that there would be some interaction between them. One candidate is the formula $R_r\varphi \to (\varphi \to [r:\varphi?]\mathbb{O}_s\varphi)$, expressing that if the receiver has the power to know whether φ and φ is the case, then the sender is obliged to announce φ once the receiver has asked the question φ?. It is not hard to show the validity of the formula when φ is propositional. But the next proposition shows that this needs not to be the case when φ is a general formula:

Proposition 3.3 *$\not\models R_r\varphi \to (\varphi \to [r:\varphi?]\mathbb{O}_s\varphi)$ for some formulas φ.*

Proof. We show that $\not\models R_r(p\wedge\neg\mathbb{O}_sp) \to ((p\wedge\neg\mathbb{O}_sp) \to [r:(p\wedge\neg\mathbb{O}_sp)?]\mathbb{O}_s(p\wedge\neg\mathbb{O}_sp))$. Let the model $M = (W, \sim, \approx, N, V)$ be as follows:

- $W = \{w, v\}$, $\sim = W \times W$, $\approx = \{\{v\}, \{w\}\}$, $V(p) = \{w\}$,
- $N(w) = \{\{w\}, \{v, w\}\}$ and $N(v) = \{v, w\}$.

We are going to show that $M, w \not\models R_r(p \wedge \neg\mathbb{O}_sp) \to ((p \wedge \neg\mathbb{O}_sp) \to [r:(p \wedge \neg\mathbb{O}_sp)?]\mathbb{O}_s(p \wedge \neg\mathbb{O}_sp))$. First, it is not hard to see that $M, w \models p \wedge \neg\mathbb{O}_sp$ (1) and $M, v \not\models p\wedge\neg\mathbb{O}_sp$. Hence $[\![p\wedge\neg\mathbb{O}_sp]\!]_M = \{w\}$ and $[\![\neg(p\wedge\neg\mathbb{O}_sp)]\!]_M = \{v\}$. Therefore $M, w \models R_r(p\wedge\neg\mathbb{O}_sp)$ (2). In the updated model $M_{(p\wedge\neg\mathbb{O}_sp)?}$, the only change is that $N_{(p\wedge\neg\mathbb{O}_sp)?}(w) = \{\{w\}\}$. It follows that $M_{(p\wedge\neg\mathbb{O}_sp)?}, w \models \mathbb{O}_sp$. Thus $M_{(p\wedge\neg\mathbb{O}_sp)?}, w \not\models \mathbb{O}_s(p\wedge\neg\mathbb{O}_sp)$. Therefore $M, w \not\models [r:(p\wedge\neg\mathbb{O}_sp)?]\mathbb{O}_s(p\wedge\neg\mathbb{O}_sp)$ (3). By (1), (2), and (3), $M, w \not\models R_r(p \wedge \neg\mathbb{O}_sp) \to ((p \wedge \neg\mathbb{O}_sp) \to [r:(p \wedge \neg\mathbb{O}_sp)?]\mathbb{O}_s(p \wedge \neg\mathbb{O}_sp))$. □

In the above proof, the formula $p\wedge\neg\mathbb{O}_sp$ is used to show the invalidity of the given axiom schema. The formula has the same structure as the Moore sentence [17], i.e., p is true but I do not believe p. It is well known, in dynamic epistemic logic [22], that the Moore sentence $(p \wedge \neg Kp)$ is an unsuccessful formula, in the sense that the Moore sentence may become false after the announcement of itself. Here we see a similar situation: after the receiver asks the question $p\wedge\neg\mathbb{O}_sp$?, it becomes obligatory for the sender to announce p. Thus the truth of the formula $p \wedge \neg\mathbb{O}_sp$ becomes false. However, the operator $\mathbb{O}_s$ satisfies a

[8] This can be seen as an analogy to the rule of necessitation in SDL.

weak form of the axiom (T): $\models \neg\mathbb{O}_s\bot \to (\mathbb{O}_s\varphi \to \varphi)$ (Proposition 3.2(2)). Hence, in the updated model, it is not the case that $\mathbb{O}_s(p \wedge \neg\mathbb{O}_s p)$.

At first glance, the invalidity of the axiom schema in Proposition 3.3 may seem to be counterintuitive. How could it be that the receiver has the power to know something while the sender has no obligation to inform even if the receiver requests it? We will, nevertheless, argue that the phenomenon can be explained if we make explicit the time involved in the axiom schema. The operator $\mathbb{O}_s\varphi$ in LRK expresses veritable obligations ([8], i.e., obligations specific to a particular situation) instead of normative rules. This means that the truth of formulas like $\mathbb{O}_s\varphi$ may flip after the receiver asks some questions because the situation changes. Consider the formula $R_r(p \wedge \neg\mathbb{O}_s p) \to ((p \wedge \neg\mathbb{O}_s p) \to [r : (p\wedge\neg\mathbb{O}_s p)?]\mathbb{O}_s(p\wedge\neg\mathbb{O}_s p))$. The first three occurrences of the operator $\mathbb{O}_s$ really refer to the obligation of the sender *before* the question $(p \wedge \neg\mathbb{O}_s p)?$, whereas the last two occurrences express the sender's obligation *after* the question. Thus, the antecedent $R_r(p\wedge\neg\mathbb{O}_s p)$ just asserts that the receiver has the power to know the sender's deontic status before the question. However, in LRK, there is no way to express the sender's previous obligation in the scope of the dynamic operator $[r : p \wedge \neg\mathbb{O}_s p?]$. This suggests that LRK may be equipped with temporal operators like "Yesterday".

The next two propositions are on the behavior of the two dynamic operators $[s{:}\varphi!]$ and $[r{:}\varphi?]$. Like in dynamic epistemic logic, a series of reduction axioms exist for these two operators. In order to show this, we need Lemma 3.4.

Lemma 3.4 *The following hold:*

(1) for any model M and formulas φ, ψ, $[\![\langle s{:}\varphi!\rangle\psi \vee \langle s{:}\neg\varphi!\rangle\psi]\!]_M = [\![\psi]\!]_{M_{\varphi!}}$.

(2) for any pointed model M, w and formulas φ, ψ, if $M, w \models R_r\varphi$ then $[\![[r{:}\varphi?]\psi]\!]_M = [\![\psi]\!]_{M_{\varphi?}}$.

Proof. The proof for (2) is straightforward. We consider only (1): let $M = (W, \sim, \approx, N, V)$ and $w \in W$. We have

$$\begin{array}{lll}
 & M, w \models \langle s{:}\varphi!\rangle\psi \vee \langle s{:}\neg\varphi!\rangle\psi & \\
\text{iff} & M, w \models \langle s{:}\varphi!\rangle\psi \text{ or } M, w \models \langle s{:}\neg\varphi!\rangle\psi & \text{(semantics)} \\
\text{iff} & (M, w \models \varphi \text{ and } M_{\varphi!}, w \models \psi) \text{ or} & \text{(semantics)} \\
 & (M, w \models \neg\varphi \text{ and } M_{\neg\varphi!}, w \models \psi) & \\
\text{iff} & (M, w \models \varphi \text{ and } M_{\varphi!}, w \models \psi) \text{ or} & (M_{\varphi!} = M_{\neg\varphi!}) \\
 & (M, w \models \neg\varphi \text{ and } M_{\varphi!}, w \models \psi) & \\
\text{iff} & M_{\varphi!}, w \models \psi &
\end{array}$$

Proposition 3.5 *The following hold for all formulas φ and ψ:*

(1) $\models [s{:}\varphi!]p \leftrightarrow (\varphi \to p)$.

(2) $\models [s{:}\varphi!]\neg\psi \leftrightarrow (\varphi \to \neg[s{:}\varphi!]\psi)$.

(3) $\models [s{:}\varphi!](\psi \to \chi) \leftrightarrow ([s{:}\varphi!]\psi \to [s{:}\varphi!]\chi)$.

(4) $\models [s{:}\varphi!]K_r\psi \leftrightarrow (\varphi \to K_r[s{:}\varphi!]\psi)$.

(5) $\models [s{:}\varphi!]R_r\psi \leftrightarrow (\varphi \to R_r(\langle s{:}\varphi!\rangle\psi \vee \langle s{:}\neg\varphi!\rangle\psi))$.

(6) $\models [s{:}\varphi!]\Box\psi \leftrightarrow (\varphi \to \Box(\langle s{:}\varphi!\rangle\psi \vee \langle s{:}\neg\varphi!\rangle\psi))$.

(7) if $\models \varphi$ *then* $\models [s\!:\!\psi!]\varphi$.

Proof. We show only (5). Let $M = (W, \sim, \approx, N, V)$ and $w \in W$. We have

$$\begin{array}{lll}
 & M, w \models [s\!:\!\varphi!]R_r\psi & \\
\text{iff} & M, w \models \varphi \text{ implies } M_{\varphi!}, w \models R_r\psi & \text{(semantics)} \\
\text{iff} & M, w \models \varphi \text{ implies} & \text{(semantics)} \\
 & (\forall U \in \approx_{\varphi!}\colon\ U \subseteq \llbracket\psi\rrbracket_{M_{\varphi!}} \text{ or } U \subseteq \llbracket\neg\psi\rrbracket_{M_{\varphi!}}) & \\
\text{iff} & M, w \models \varphi \text{ implies} & \text{(def. of } M_{\varphi!}) \\
 & (\forall U \in \approx\colon\ U \subseteq \llbracket\psi\rrbracket_{M_{\varphi!}} \text{ or } U \subseteq \llbracket\neg\psi\rrbracket_{M_{\varphi!}}) & \\
\text{iff} & M, w \models \varphi \text{ implies} & \text{(Lemma 3.4(1))} \\
 & (\forall U \in \approx\colon\ (U \subseteq \llbracket\langle s\!:\!\varphi!\rangle\psi \vee \langle s\!:\!\neg\varphi!\rangle\psi\rrbracket_M \text{ or} & \\
 & U \subseteq \llbracket\neg(\langle s\!:\!\varphi!\rangle\psi \vee \langle s\!:\!\neg\varphi!\rangle\psi)\rrbracket_M)) & \\
\text{iff} & M, w \models \varphi \text{ implies} & \text{(semantics)} \\
 & M, w \models R_r(\langle s\!:\!\varphi!\rangle\psi \vee \langle s\!:\!\neg\varphi!\rangle\psi) & \\
\text{iff} & M, w \models \varphi \to R_r(\langle s\!:\!\varphi!\rangle\psi \vee \langle s\!:\!\neg\varphi!\rangle\psi) & \text{(semantics)}
\end{array}$$

Proposition 3.6 *The following hold for all formulas* φ *and* ψ*:*

(1) $\models [r\!:\!\varphi?]p \leftrightarrow p$.

(2) $\models [r\!:\!\varphi?]\neg\psi \leftrightarrow ((\neg R_r\varphi \wedge \neg\psi) \vee (R_r\varphi \wedge \neg[r\!:\!\varphi?]\psi))$.

(3) $\models [r\!:\!\varphi?](\psi \to \chi) \leftrightarrow ([r\!:\!\varphi?]\psi \to [r\!:\!\varphi?]\chi)$.

(4) $\models [r\!:\!\varphi?]K_r\psi \leftrightarrow ((K_r\psi \wedge \neg R_r\varphi) \vee (K_r[r\!:\!\varphi?]\psi \wedge R_r\varphi))$.

(5) $\models [r\!:\!\varphi?]R_r\psi \leftrightarrow ((R_r\psi \wedge \neg R_r\varphi) \vee (R_r[r\!:\!\varphi?]\psi \wedge R_r\varphi))$.

(6) $\models [r\!:\!\varphi?]\Box\psi \leftrightarrow ((\Box\psi \wedge \neg R_r\varphi) \vee (\Box[r\!:\!\varphi?]\psi \wedge R_r\varphi))$.

Proof. We show only (2) and (4). Let $M = (W, \sim, \approx, N, V)$ and $w \in W$. (2): It is clear that $M, w \models [r\!:\!\varphi?]\neg\psi \wedge \neg R_r\varphi$ iff $M, w \models \neg\psi \wedge \neg R_r\varphi$ $(*)$. On the other hand, we have:

$$\begin{array}{lll}
 & M, w \models [r\!:\!\varphi?]\neg\psi \wedge R_r\varphi & \\
\text{iff} & M_{\varphi?}, w \models \neg\psi \text{ and } M, w \models R_r\varphi & \text{(semantics)} \\
\text{iff} & M_{\varphi?}, w \not\models \psi \text{ and } M, w \models R_r\varphi & \text{(semantics)} \\
\text{iff} & M, w \not\models [r\!:\!\varphi?]\psi \text{ and } M, w \models R_r\varphi & \text{(semantics)} \\
\text{iff} & M, w \models \neg[r\!:\!\varphi?]\psi \wedge R_r\varphi & \text{(semantics)}
\end{array}$$

Hence $\models ([r\!:\!\varphi?]\neg\psi \wedge R_r\varphi) \leftrightarrow (\neg[r\!:\!\varphi?]\psi \wedge R_r\varphi)$ $(**)$. From $(*)$ and $(**)$ it follows that $\models [r\!:\!\varphi?]\neg\psi \leftrightarrow ((\neg R_r\varphi \wedge \neg\psi) \vee (R_r\varphi \wedge \neg[r\!:\!\varphi?]\psi))$.

(4): It is clear that $\models ([r\!:\!\varphi?]K_r\psi \wedge \neg R_r\varphi) \leftrightarrow (K_r\psi \wedge \neg R_r\varphi)$ $(*)$. On the other hand:

$$\begin{array}{ll}
 & M, w \models [r\!:\!\varphi?]K_r\psi \wedge R_r\varphi \\
\text{iff} & M_{\varphi?}, w \models K_r\psi \text{ and } M, w \models R_r\varphi \text{ (semantics)} \\
\text{iff} & (\forall v \in W_{\varphi?}\colon\ w \sim_{\varphi?} v \text{ implies } M_{\varphi?}, v \models \psi) \text{ and } M, w \models R_r\varphi \text{ (semantics)} \\
\text{iff} & (\forall v \in W\colon\ w \sim v \text{ implies } M_{\varphi?}, v \models \psi) \text{ and} M, w \models R_r\varphi \text{ (def. of } M_{\varphi?}) \\
\text{iff} & (\forall v \in W\colon\ w \sim v \text{ implies } M, v \models [r\!:\!\varphi?]\psi) \text{ and } M, w \models R_r\varphi \\
 & \text{(Lemma 3.4(2))} \\
\text{iff} & M, w \models K_r[r\!:\!\varphi?]\psi \wedge R_r\varphi \text{ (semantics)}
\end{array}$$

Therefore $\models ([r : \varphi?]K_r\psi \wedge R_r\varphi) \leftrightarrow (K_r[r : \varphi?]\psi \wedge R_r\varphi)$ $(**)$. From $(*)$ and

($**$) it follows that $\models [r\!:\!\varphi?]K_r\psi \leftrightarrow ((K_r\psi \wedge \neg R_r\varphi) \vee (K_r[r\!:\!\varphi?]\psi \wedge R_r\varphi))$. □

4 Expressivity

In this section, we investigate the expressive power of some fragments of $\mathcal{L}$. As we have seen, there are a couple of operators in $\mathcal{L}$. But it is not clear whether some of them are superfluous. In particular, we are going to address the following two questions in the section:

- Does adding the two dynamic operator $[s\!:\!\varphi!]$ and $[r:\varphi?]$ to $\mathcal{L}$ increases the expressive power?
- Is the operator $R_r\varphi$ for the power to know expressible in a language without it?

It is well known that adding the public announcement operator to the language of epistemic logic (without common knowledge) does not increase expressive power. In Section 3, we have seen that there exists a series of reduction axioms for the two dynamic operators $[s\!:\!\varphi!]$ and $[r\!:\!\varphi?]$, but not for formulas of the form $[s\!:\!\varphi!]\mathbb{O}_s\psi$ and $[r\!:\!\varphi?]\mathbb{O}_s\psi$. Thus it is interesting to know whether the same holds for LRK. To answer the first question, let $\mathcal{L}_0$, $\mathcal{L}_1$, and $\mathcal{L}_2$ be the sublanguages of $\mathcal{L}$ defined as follows:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \to \varphi) \mid K_r\varphi \mid R_r\varphi \mid \mathbb{O}_s\varphi \mid \Box\varphi \qquad (\mathcal{L}_0)$$

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \to \varphi) \mid K_r\varphi \mid R_r\varphi \mid \mathbb{O}_s\varphi \mid \Box\varphi \mid [s\!:\!\varphi!]\varphi \qquad (\mathcal{L}_1)$$

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \to \varphi) \mid K_r\varphi \mid R_r\varphi \mid \mathbb{O}_s\varphi \mid \Box\varphi \mid [r\!:\!\varphi?]\varphi \qquad (\mathcal{L}_2)$$

The next theorem shows that adding any of the two operators $[s:\varphi!]\psi$ and $[r:\varphi?]\psi$ to the static language $\mathcal{L}_0$ does increase the expressive power. This is in contrast with the situation in public announcement logic and gives a positive answer to the first question. It also follows that no DEL-style reduction axiomatization exists for LRK.

Theorem 4.1 *The following hold:*

(1) $\mathcal{L}_1$ is more expressive than $\mathcal{L}_0$.

(2) $\mathcal{L}_2$ is more expressive than $\mathcal{L}_0$.

Proof. (1): It is clear that $\mathcal{L}_1$ is at least as expressive as $\mathcal{L}_0$ since $\mathcal{L}_0$ is a sublanguage of $\mathcal{L}_1$. We show that $\mathcal{L}_0$ is not at least as expressive as $\mathcal{L}_1$. This is done by showing that there is no $\psi \in \mathcal{L}_0$ such that $[s:p!]\mathbb{O}_s\bot \equiv \psi$ (i.e., they are satisfied at exactly the same pointed models). Consider two models $M_1 = (W, \sim, \approx, N_1, V)$ and $M_2 = (W, \sim, \approx, N_2, V)$ where:

- $W = \{w, u\}$, $\sim = W \times W$, $\approx = \{\{w\}, \{u\}\}$, $V(p) = \{w\}$;
- $N_1(w) = \{\{w, u\}\}$, $N_2(w) = \{\{w\}, \{w, u\}\}$, $N_1(u) = N_2(u) = \{W\}$.

It is not hard to see that $M_1, w \models [s\!:\!p!]\mathbb{O}_s\bot$ and $M_2, w \not\models [s\!:\!p!]\mathbb{O}_s\bot$. However, by an induction on the structure of ψ, we can show that $M_1, y \models \psi$ iff $M_2, y \models \psi$ for all $\psi \in \mathcal{L}_0$ and $y \in W$. Here we show only the inductive step for $\mathbb{O}_s\chi$. The case for $y = u$ follows directly from that $N_1(u) = N_2(u)$ and the IH. For $y = w$, we have:

$$\begin{array}{lll}
 & M_1, w \models \mathbb{O}_s\chi & \\
\text{iff} & \forall U \in N_1(w),\ U \subseteq\sim (w) \text{ implies } U \subseteq [\![\chi]\!]_{M_1} & \text{(semantics)} \\
\text{iff} & \forall U \in N_1(w),\ U \subseteq [\![\chi]\!]_{M_1} & \text{(def. of } N_1(w) \text{ and } \sim) \\
\text{iff} & \forall U \in N_1(w),\ U \subseteq [\![\chi]\!]_{M_2} & \text{(IH)} \\
\text{iff} & \forall U \in N_2(w),\ U \subseteq [\![\chi]\!]_{M_2} & (\bigcup N_1(w) = \bigcup N_2(w)) \\
\text{iff} & \forall U \in N_2(w),\ U \subseteq\sim (w) \text{ implies } U \subseteq [\![\chi]\!]_{M_2} & \text{(def. of } N_2(w) \text{ and } \sim) \\
\text{iff} & M_2, w \models \mathbb{O}_s\chi & \text{(semantics)}
\end{array}$$

(2): (2) can be shown similarly to (1). To obtain proof, we replace the formula $[s\!:\!p!]\mathbb{O}_s\bot$ in the proof of (1) by $[r\!:\!p?]\mathbb{O}_s\bot$. □

The remainder of the section is devoted to the second question. We want to know whether the notion of the power to know characterized in LRK is reducible to (a combination of) other operators. To do this, we consider the following sublanguage of $\mathcal{L}$ without the operator $R_r\varphi$:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \to \varphi) \mid K_r\varphi \mid \mathbb{O}_s\varphi \mid \Box\varphi \mid [s\!:\!\varphi!]\varphi \mid [r\!:\!\varphi?]\varphi \qquad (\mathcal{L}_3)$$

The usual way to proceed is to find a suitable notion of "bisimulation" for $\mathcal{L}_3$. In the next definition, we define the "almost-identical" relation between two models. Although the notion is much stronger than the usual notion of bisimulation, it is sufficient for our current purpose.

Definition 4.2 Let $M = (W, \sim, \approx, N, V)$ and $M' = (W', \sim', \approx', N', V')$ be two models. We say M and M' are *almost-identical* if the following hold:

(A1) $W = W'$, $\sim = \sim'$, $N = N'$, $V = V'$, and

(A2) for all $w \in W$ and $U \in N(w)$, $U \subseteq\approx (w)$ and $U \subseteq\approx' (w)$.

Lemma 4.3 *For all models $M = (W, \sim, \approx, N, V)$ and $M' = (W', \sim', \approx', N', V')$, if M and M' are almost-identical then $M, w \models \varphi$ iff $M', w \models \varphi$ for all $w \in W$ and $\varphi \in \mathcal{L}_3$.*

Proof. Induction on the structure of φ. We show only the cases for $[s\!:\!\psi!]\chi$ and $[r\!:\!\psi?]\chi$ since the remaining are all straightforward.

Case $[s\!:\!\psi!]\chi$: we have

$$\begin{array}{lll}
 & M, w \models [s\!:\!\psi!]\chi & \\
\text{iff} & M, w \not\models \psi \text{ or } M_{\psi!}, w \models \chi & \text{(semantics)} \\
\text{iff} & M', w \not\models \psi \text{ or } M_{\psi!}, w \models \chi & \text{(IH)} \\
\text{iff} & M', w \not\models \psi \text{ or } M'_{\psi!}, w \models \chi & \text{(IH)} \\
\text{iff} & M', w \models [s\!:\!\psi!]\chi & \text{(semantics)}
\end{array}$$

Note that the third "iff" holds because $M_{\psi!}$ and $M'_{\psi!}$ are almost-identical since $[\![\psi]\!]_M = [\![\psi]\!]_{M'}$ by the IH.

Case $[r\!:\!\psi?]\chi$: we have[9]

[9] We need to slightly modify the definition of $M_{\psi?}$ (and $M'_{\psi?}$) below since the operator $R_r\psi$

$$\begin{array}{lll}
 & M, w \models [r\!:\!\psi?]\chi & \\
\text{iff} & M_{\psi?}, w \models \chi & \text{(semantics)} \\
\text{iff} & M, w \models \chi & (M_{\psi?} = M \text{ by the condition (A2)}) \\
\text{iff} & M', w \models \chi & \text{(IH)} \\
\text{iff} & M'_{\psi?}, w \models \chi & (M'_{\psi?} = M' \text{ by the condition (A2)}) \\
\text{iff} & M', w \models [r\!:\!\psi?]\chi & \text{(semantics)}
\end{array}$$

The next proposition gives a negative answer to the second question. The operator $R_r\varphi$ does express an independent notion.

Proposition 4.4 *There is no formula $\varphi \in \mathcal{L}_3$ such that $\varphi \equiv R_r p$.*

Proof. Consider two models $M = (W, \sim, \approx, N, V)$ and $M' = (W, \sim, \approx', N, V)$ such that $W = \{w, v\}$, $\sim = W \times W$, $N(w) = \{\{w\}\}$, $N(v) = \{\{v\}\}$, $V(p) = \{w\}$, and

- $\approx = \{\{w\}, \{v\}\}$,
- $\approx' = \{\{w, v\}\}$.

It is easy to see that $M, w \models R_r p$ and $M', w \not\models R_r p$. However, note that M and M' are almost-identical. Hence, $M, w \models \varphi$ iff $M', w \models \varphi$ for all $\varphi \in \mathcal{L}_3$ by Lemma 4.3. □

Corollary 4.5 *$\mathcal{L}$ is more expressive than $\mathcal{L}_3$.*

5 Related Work

Deontic logic for epistemic actions. In general, LRK is a specific deontic logic for actions in the epistemic context. Several attempts at developing deontic logic for epistemic actions can be found in the literature, e.g., [2], [1], and [12].

A logic for the notion of permitted announcements has been developed in [12]. The logic is based on the idea that a piece of information φ is permitted to be announced ($\mathbb{P}\varphi$) if the *epistemic state* after the announcement is ideal. The notion of epistemic state can be understood either as a syntactic notion or as a semantic notion. Syntactically, an epistemic state is just a set of epistemic formulas representing the knowledge of an agent. In contrast, from the semantic perspective, an epistemic state is a set of indistinguishable possible worlds by an agent (or, equivalently, an indistinguishability relation over all possible worlds). In [12], two semantic definitions of permitted announcements have been proposed based on the two understandings of epistemic states. Interestingly, the two definitions are shown to be equivalent, in the sense that they give the same set of logical validities. In this paper, we consider epistemic states as a semantic notion and apply the "neighbourhood epistemic model" introduced in [12]. Formally, a neighbourhood epistemic model is a structure $M = (W, \sim, N, V)$ where W, $\sim$, and V are the same as in the standard epistemic models (S5 models) and $N : W \to \wp(\wp(W))$ is a neighbourhood function assigning a set of *ideal* epistemic states to each possible world. Note that the

does not appear in $\mathcal{L}_3$. For example, $M_{\psi?}$ should be defined as "$M_{\psi?} = M$ if it is not the case that for all $U \in \approx$, $U \subseteq [\![\psi]\!]_M$ or $U \subseteq [\![\neg\psi]\!]_M$; otherwise ...".

public announcement of a proposition φ may restrict the current epistemic state to only those possible worlds satisfying φ. Hence, in [12], the formula $\mathbb{P}\varphi$ is interpreted in such a way that it is true at a possible world w iff the epistemic state after the announcement of φ is an element of $N(w)$. In this paper, we employ a similar idea to provide semantics for the notion of obligatory announcements.

In [2], two binary operators $P(\psi, \varphi)$ and $O(\psi, \varphi)$ are introduced to express the notions that "after announcing ψ, it is permitted/obligatory to announce φ". It is clear that our operator $\mathbb{O}_s\varphi$ can be expressed in their framework as $O(\top, \varphi)$. Conversely, the operator $O(\psi, \varphi)$ can be expressed as $[s : \psi!]\mathbb{O}_s\varphi$ in LRK. In [2], a ternary relation $\mathcal{P} \subseteq S \times \wp(S) \times \wp(S)$ is used to provide the semantics for $O(\psi, \varphi)$ in such a way that $\mathcal{M}, s \models O(\psi, \varphi)$ iff for all $(s, [\![\psi]\!]_M, S'') \in \mathcal{P}, S'' \subseteq [\![\langle\psi\rangle\varphi]\!]_M$, where S is the domain of the model. The major difference between the semantics for $\mathbb{O}_s\varphi$ and $O(\psi, \varphi)$ is that our operator $\mathbb{O}_s\varphi$ is specific to the receiver's knowledge. This is reflected in the fact that the formula $K_r(\varphi \to \psi) \to (\mathbb{O}_s\varphi \to \mathbb{O}_s\psi)$ is valid in LRK whereas not in the logic of [2]. We think that LRK is more suitable for reasoning about obligatory announcements in the context of, e.g., database security, because the receiver's initial knowledge is crucial for the sender's decision on which information should be disclosed, as suggested by Example 2.7.

An alternative definition of obligatory announcements has also been proposed in [1, Definition 10]. Aucher et al. [1] define the obligatory message of a security monitor as the minimal informative message such that, by sending it, the privacy policy compliance is restored. Clearly, the notion of "obligatory message" is different from the notion of obligatory announcements in our paper and [2]. But a detailed conceptual analysis of the difference between these notions is beyond the scope of the paper.

Logic of questions. The semantics of questions or interrogatives has received much attention in logic, see [7]. The basic idea is that the meaning of a question is what counts as an answer to that question. For example, the question "Is it raining in Guangzhou?" has two possible answers: "It is raining in Guangzhou" and "It is not raining in Guangzhou". Observe that they are both propositions or statements and, furthermore, they logically exclude each other and are jointly exhaustive. Hence, some logicians propose that questions can be represented semantically as a partition over the set of all possible worlds (or the logical space), e.g., [7], [4], and [21]. The semantics for questions proposed in [7], [4], and [21] are different. In this paper, we follow the approach in [21], because it is a conservative extension to the standard epistemic logic and we want to focus on the semantics for the power to know.

In [21], the so-called *epistemic issue models* are used to provide the semantics for questions. Formally, they are structures $M = (W, \sim, \approx, V)$ where the only novel thing is the equivalence relation $\approx$ on W (or, equivalently, a partition of W). Instead of having a single modality expressing that "the question whether φ is one of the current issues", a complex formula is used in [21] to express the notion. Technically, in addition to the modality $K\varphi$ for knowledge, there are two new modalities $U\varphi$ and $Q\varphi$ where $U\varphi$ is the familiar universal modality and the truth definition of $Q\varphi$ is as follows:

$$M, w \models Q\varphi \text{ iff for all } v \in W\colon\ w \approx v \text{ implies } M, v \models \varphi$$

The notion that "the question whether φ is one of the current issues" is then expressed as the formula $U(Q\varphi \vee Q\neg\varphi)$.

As mentioned above, the idea behind the logic of question in [21] is that a set of questions can be represented as a partition over the logical space. But the logic remains neutral about where the partition or the set of questions is induced. For example, it can be induced either by a conversation, or by a game, or even by a research program [21]. In this paper, we assume that the partition is induced by the part of a normative system, such as a privacy policy, stipulating the questions an agent has the power to know the answers to.

Our work is closely related to [21] as the fragment of LRK on the operator $R_r\varphi$ is roughly a reinterpretation of the static logic $\mathrm{EL_Q}$ in [21]. However, there is also an important difference between our paper and [21]. In [21], there is also a dynamic operator $[\varphi?]\psi$ expressing that ψ holds after asking the question whether φ. But the effect of $[\varphi?]$ is to add the question $\varphi?$ to the set of current issues. In contrast, the operator $[r:\varphi?]\psi$ in LRK captures the deontic aspect of asking questions. It seems more appropriate to interpret the operator $[r:\varphi?]$ in our paper as the action that *commands* the sender to inform whether φ.

Logic of legal competence. The right to know is an epistemic right, and thus is a form of right. Works on the logical analysis of legal rights can also be found in the literature, e.g., [11,13,14,10]. Recent works on the topic investigating explicitly the power type of right includes [15,19,6]. Given that these works are on general (power-)rights, one may wonder why there is a need to develop a separate logic for epistemic rights. One reason is that there are some valid reasoning patterns for epistemic rights that can not be expressed in a language devised for general rights, just like not all principles of public announcement logic can be expressed in dynamic logic.

In our paper, we treat the power to know as an independent notion. As pointed out by [6], there exist two different logical approaches formalizing legal power: an earlier tradition reduces power to (a combination of) obligations, permissions, and actions ([11,13]); whereas the other (e.g.[14,10]) holds that power is not reducible to static normative positions, which follows the original separation of Hohfeld (see [15]). Thus, our work adopts the second approach. We also show, in Proposition 4.4, that the notion of the power to know characterized in LRK can not be expressed in a language without it.

6 Conclusion and Future Work

In this paper, a logic LRK was introduced semantically for reasoning about the power to know, the obligatory announcements, and the dynamics of questions and public announcements. We explored some (in)validities of LRK, where the interaction between the power to know and the obligatory announcements has been highlighted. We also studied the expressive power of several fragments of the language of LRK. We showed that the incorporation of the two dynamic operators in LRK increases the expressive power and the operator for the power to know cannot be expressed in a language without it.

There are many directions for future research. A natural task is to investigate some metalogical properties of LRK, such as axiomatization, completeness, and computational complexity. We can also consider extensions to LRK. For example, it is interesting to reason about the ability of the receiver in LRK since the receiver may use their power to know by asking (a sequence of) questions. Other interesting extensions to LRK include the incorporation of "the power not to know" [16], how to model the actions of adding or removing the receiver's power to know, how to extend to the multi-agent case, etc.

References

[1] Aucher, G., G. Boella and L. van der Torre, *A dynamic logic for privacy compliance*, Artificial Intelligence and Law **19** (2011), p. 187.

[2] Balbiani, P. and P. Seban, *Reasoning about permitted announcements*, Journal of Philosophical Logic **40** (2011), pp. 445–472.

[3] Bonatti, P., S. Kraus and V. Subrahmanian, *Foundations of secure deductive databases*, IEEE Transactions on Knowledge and Data Engineering **7** (1995), pp. 406–422.

[4] Ciardelli, I. A., "Questions in logic," Ph.D. thesis, University of Amsterdam (2015).

[5] Cuppens, F. and R. Demolombe, *A deontic logic for reasoning about confidentiality*, in: M. A. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems* (1996), pp. 66–79.

[6] Dong, H. and O. Roy, *Dynamic logic of legal competences*, Journal of Logic, Language and Information **30** (2021), pp. 701–724.

[7] Groenendijk, J. and M. Stokhof, *Chapter 19 - questions*, in: J. van Benthem and A. ter Meulen, editors, *Handbook of Logic and Language*, North-Holland, Amsterdam, 1997 pp. 1055–1124.

[8] Hansson, S. O., *Alternative semantics for deontic logic*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, College Publications, 2013 pp. 445–497.

[9] Hohfeld, W. N., *Fundamental legal conceptions applied in judicial reasoning*, in: W. W. Cook, editor, *Fundamental Legal Conceptions Applied in Judicial Reasoning and Other Legal Essays*, New Haven: Yale University Press, 1923 pp. 23–64.

[10] Jones, A. J. I. and M. Sergot, *A Formal Characterisation of Institutionalised Power*, Logic Journal of the IGPL **4** (1996), pp. 427–443.

[11] Kanger, S. and H. Kanger, *Rights and parliamentarism*, Theoria **32** (1966), pp. 85–115.

[12] Li, X., D. Gabbay and R. Markovich, *Dynamic Deontic Logic for Permitted Announcements*, in: *Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning*, 2022, pp. 226–235.

[13] Lindahl, L., "Position and change: A study in law and logic," Synthese Library, Springer Dordrecht, 1977, 1 edition.

[14] Makinson, D., *On the formal representation of rights relations*, Journal of Philosophical Logic **15** (1986), pp. 403–425.

[15] Markovich, R., *Understanding Hohfeld and Formalizing Legal Rights: the Hohfeldian Conceptions and Their Conditional Consequences*, Studia Logica **108** (2020).

[16] Markovich, R. and O. Roy, *Formalizing the right to know: Epistemic rights as normative positions*, in: *The First International Workshop on Logics for New-Generation Artificial Intelligence (LNGAI 2021)*, 2021, pp. 154–159.

[17] Moore, G. E., *A reply to my critics*, in: P. A. Schilpp, editor, *The Philosophy of G.E. Moore*, The Library of Living Philosophers **4**, Northwestern University, Evanston, Illinois, 1942 pp. 535–677.

[18] Sergot, M., *Normative Positions*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, College Publications, 2013 pp. 353–406.

[19] Sileno, G. and M. Pascucci, *Disentangling deontic positions and abilities: a modal analysis*, in: F. Calimeri, S. Perri and E. Zumpano, editors, *Proceedings of the 35th Edition of the Italian Conference on Computational Logic (CILC 2020)*, 2020, pp. 36–50.

[20] van Benthem, J. and F. Liu, *Dynamic logic of preference upgrade*, Journal of Applied Non-Classical Logics **17** (2007), pp. 157–182.

[21] van Benthem, J. and Ş. Minică, *Toward a dynamic logic of questions*, Journal of Philosophical Logic **41** (2012), pp. 633–669.

[22] van Ditmarsch, H., W. van der Hoek and B. Kooi, "Dynamic Epistemic Logic," Springer, Dordrecht, 2008.

[23] Watson, L., "The Right to Know: Epistemic Rights and Why We Need Them," Routledge Focus on Philosophy, Routledge, 2021, 1 edition.

Perspectival obligation and extensionality in an alethic-deontic setting

Dominik Pichler and Xavier Parent [1]

Favoritenstraße 9/E192-05 (Stiege 2/3.Stock) A-1040 Wien, Austria

Abstract

The theme of extensionality in first-order deontic logic has been thoroughly studied in the past, but not in the context of a combination of different types of modalities. An operator is extensional if it allows substitution *salva veritate* of co-referential terms within its scope and intensional if it does not. It can be argued that one distinctive feature of "ought" (as opposed to the other modalities) is that it is extensional. The question naturally arises as to whether it is possible to combine extensionality and intensionality of different modal operators in the same semantics without creating the deontic collapse. We answer this question within a particular framework, Åqvist's system **F** for conditional obligation. We develop in full detail a perspectival account of obligation (and related notions), as was done for Standard Deontic Logic (SDL) by Goble. It is called "perspectival", because one always evaluates the content of an obligation in one world from the perspective of another one, hence using some form of cross-world evaluation. The proposed framework allows for a more nuanced way of approaching first-order deontic principles.

Keywords: First-order reasoning, extensionality, conditional obligation, 2-dimensional semantics, preferences, perspectivism

1 Introduction

The past 15 years have seen a renewed interest in so-called relativism or perspectivism in the philosophy of language. Relativist or perspectivist accounts have been put forth to explain discourse about knowledge, epistemic possibility, matters of taste, contingent future events, modalities (including the deontic ones) and the like. Here relativism is usually taken to be, or to presuppose, a semantic thesis. Understanding how some discourses function requires recognizing that speakers express propositions whose truth or falsity are relative to parameters or perspectives in addition to a possible world−see Kölbel [20] for a

[1] Emails: dominik@logic.at; xavier@logic.at. The research of the second author is supported by the Austrian Science Fund (FWF) through the ANCoR project [M-3240-N]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. This paper was originated by the first author's master thesis [25]. The material was modified, and expanded. We thank three anonymous reviewers for their valuable comments.

thorough defense of this view, and also MacFarlane [22]. The approach is often called "perspectivism" as it has a less negative connotation than "relativism", and we will stick to this term.
The purpose of the present paper is to show some of the usefulness of this view for normative reasoning. We believe it may shed light on a topic that has been overlooked in the recent papers devoted to first-order deontic reasoning, e.g. [7,8,28]. This is the topic of extensionality of "ought". We do not claim to be original, as we will pick up on a proposal made long ago by Goble [12,13,14]. It can be summarized thus. An operator is extensional if it allows substitution *salva veritate* of co-referential terms within its scope, and intensional if it does not. It can be argued that one distinctive feature of "ought" (as opposed to the other modalities) is that it is extensional. The problem is: a deontic logic in which "ought" is extensional can be shown to collapse to triviality. Goble developed his own solution to this problem, and we will refer to it as the original "perspectival" account. The basic idea is that the content of an obligation at one world is to be evaluated from the perspective of another one, so that some form of cross-world evaluation is made possible. This idea of cross-world evaluation is familiar from the literature on multi-dimensional modal logic (see e.g. [3,11,18,29]). Other works in multi-dimensional deontic logic we are aware of focus on the propositional case [6,9,10,17]. The novelty lies in linking the perspectival idea to first-order considerations.

Our goal is to improve the original account in two ways. By doing so, we hope to strengthen the case for the perspectival idea, and provide more credibility to it.

- The original account is cast within the framework of Standard Deontic Logic (SDL) [31], which is known to be plagued by the deontic paradoxes, in particular the paradox of contrary-to-duty (CTD) obligation [4]. We will recast the account within the framework of preference-based dyadic deontic logic [1,5,15,16,23]. Dyadic deontic logic is the logic for reasoning with dyadic obligations "it ought to be the case that ψ if it is the case that φ" (notation: $\bigcirc(\psi/\varphi)$). Its semantics is in terms of a betterness relation. Initially devised to resolve the CTD paradox, dyadic deontic logic is a recognized standard for normative reasoning. The idea of making it two-dimensional is not entirely new: Lewis [21, p. 63] suggested to analyze conditionals within the framework of two-dimensional modal logic, but his motivations were different.
- The original account does not allow for different types of modalities to interact. We will lift this restriction, and look at the question of whether it is possible to combine extensionality and intensionality of different modal operators in the same semantics without creating the collapse. We will use Åqvist's mixed alethic-deontic preference-based logic **F** [1,23,24]. The language of **F** has an extra modal operator $\Box$ ("it is settled that"), allowing to capture some fundamental principles of normative reasoning, like "strong factual detachment" [26]. Among the systems proposed by Åqvist, **F** is also the weakest one in which the collapse arises. The first-order ex-

tension of **F** will be called $\mathbf{F}^{\forall}$. (One could object that, in **F**, $\Box$ is a *soi disant* modality, definable in terms of $\bigcirc(-/-)$. In $\mathbf{F}^{\forall}$, it will become a first-class citizen, viz. a primitive modality.)

The paper is organized as follows. Section 2 sets the stage, and defines a list of basic requirements to be met by the logic. Section 3 develops in full semantic detail the perspectival account of obligation (and related notions) alluded to above. Section 4 shows how the requirements are met. Section 5 concludes.

2 Setting the stage

We give a list of basic requirements that we think an adequate first-order (FO) deontic logic should meet. The problem dealt with in this paper will be to devise a framework meeting them. For ease of readability, we formulate the requirements within the language of a monadic deontic logic. Our list is not meant to be exhaustive.

2.1 Requirements

Requirement 1 (Extentionality for "ought") $\bigcirc$ *("It ought to be the case that ...") should validate the principle of substitution salva veritate (E-*$\bigcirc$*), where* φ *is a formula,* t *and* s *are terms, and* $\varphi_{t\hookrightarrow s}$ *is the result of replacing zero up to all occurrences of* t*, in* φ*, by* s*:*

$$t = s \to (\bigcirc\varphi \leftrightarrow \bigcirc\varphi_{t\hookrightarrow s}) \qquad \text{(E-}\bigcirc\text{)}$$

Intuitively: two co-referential terms may be interchanged without altering the truth-value of the deontic formula in which they occur.

A modal operator is usually said to be referentially transparent, when it satisfies the principle of substitution *salva veritate*, and referentially opaque otherwise. As pointed out by Castañeda [2] there are good reasons to believe that deontic operators are referentially transparent. For instance, the inference from (1) and (2) to (3) is intuitively valid:

(1) The Pope ought to live a life of exceptional sanctity: $\bigcirc S(\imath x Po(x))$

(2) Jose is the Pope: $j = \imath x Po(x)$

(3) Jose ought to live a life of exceptional sanctity: $\bigcirc S(j)$

$\imath x Po(x)$ is a so-called definite description, and is read "the x that is Po" ("the Pope"). Definite descriptions are used to refer to what a speaker wishes to talk about. Castañeda (rightly) says: "a man's obligations are *his* [the author's emphasis] regardless of his characterizations". In other worlds, they are independent of the way he is referred to.

In daily conversations, one casually switches between a proper name and the definite description associated with it. When using one instead of the other, we are still talking about the same individual, referring to him using different descriptions (the Pope, the direct successor of St Peter, ...). This would just not be possible if "ought" was not referentially transparent.

However, it may be questioned whether the inference from (1) and (2) to (3) is valid intuitively.[2] For one can consistently add to the premises set

(2′) Jose ought not to be the Pope: $\bigcirc(j \neq \imath x Po(x))$

Does (3) still follow? It seems not. Two comments are in order. First, it may be thought that a finer-grained version of the principle is needed. To add

(2″) Jose ought to be the Pope: $\bigcirc(j = \imath x Po(x))$

would resolve the problem, but would make (2) superfluous. For (3) follows from (1) and (2″) using the standard principles of deontic logic and first-order logic. To add

(2⋆) Jose may be the Pope: $P(j = \imath x Po(x))$–P: "it is permitted that"

would resolve the problem, and not make (2) superfluous. Thus, one way to address the above problem is to introduce the following permitted version of (E-$\bigcirc$):

$$t = s \wedge P(t = s) \rightarrow (\bigcirc\varphi \leftrightarrow \bigcirc\varphi_{t\hookrightarrow s}) \qquad \text{(Permitted E-}\bigcirc\text{)}$$

Second, it may make a difference whether the substitution is done in the consequent or the antecedent of a conditional obligation. Consider:

(4) If the Pope does not live a life of exceptional sanctity, we should elect a new one: $\bigcirc(\exists y(El(y) \wedge y \neq \imath x Po(x))/\neg S(\imath x Po(x)))$

(5) If Jose does not live a life of exceptional sanctity, we should elect a new Pope: $\bigcirc(\exists y(El(y) \wedge y \neq \imath x Po(x))/\neg S(j))$

The antecedent of (4) refers to a sub-ideal world where (1) is violated. Intuitively, (4) and (5) seem equivalent, even in the presence of (2′). Thus, (4) and (5) are two different ways to say the same thing. If "ought" is not referentially transparent, then (4) and (5) are not synonymous, since they have a different antecedent. If so, one would need

- the permitted version of the principle for any substitution done in the consequent (proviso: $P(t = s)$);
- the unrestricted version for any substitution done in an antecedent.

We leave it as a topic of future research to investigate how to implement these suggestions.

For simplicity's sake, $\Box$ will be read as "It is necessary that ...". Whether it is historical necessity or some other type of necessity is not germane for our discussion.

Requirement 2 (Intensionality for "necessarily") $\Box$ *should not validate the principle of substitution salva veritate, where* t *and* s *are terms (either a constant or a definite description):*

$$t = s \rightarrow (\Box\varphi \leftrightarrow \Box\varphi_{t\hookrightarrow s}) \qquad \text{(E-}\Box\text{)}$$

[2] We owe this objection from an anonymous referee.

It is usually thought that $\Box$ should not verify (E-$\Box$). The reason why is best illustrated with the following well-known example. Intuitively, (6) and (7) do not imply (8):[3]

(6) Number of planets $= 8$

(7) $\Box(8 = 8)$

(8) $\Box$(Number of planets $= 8$)

If $\Box$ means "settled" in the sense of outside of the agent's control, then (8) is fine. But if $\Box$ means metaphysical necessity, settledness in the sense of historical necessity, or knowledge, then (8) is clearly unwanted. Indeed, before 2006, (6) was false.[4]

A second, independent argument against (E-$\Box$) will be given in Prop. 2.2.

Requirement 3 (No collapse) *The logic should avoid the deontic collapse. That is, the formula $\varphi \leftrightarrow \bigcirc\varphi$ should not be derivable.*

This requirement is taken from Goble [12,13,14]. A separate section is devoted to this requirement.

The *raison d'être* of our last requirement is this: obligations are there to make the world a better place; they are constantly violated, but should not be so. Therefore, our account should make the notion of definite description well-behaved with respect to negation. That is to say:

Requirement 4 (Self-negation) *Given E-$\bigcirc$, the logic should be able to account for the meaningfulness of a deontic statement denying a property of an individual identified using that very same property.*

Here is an example:

(9) The tyrant has an obligation not to be a tyrant: $\bigcirc\neg T(\imath x T(x))$

Self-negation like the one in (9) cannot be accounted for in (a straightforward FO extension of) SDL. (9) tells us that in the best of all possible worlds the tyrant x is not a tyrant. But this is a contradiction (assuming that such an x exists). Of course, the claim is not that in the best of all possible worlds the tyrant x is not a tyrant. Rather−to anticipate our solution−the claim is that the individual x that is a tyrant in the actual world is not a tyrant in all the best worlds. This is a relation among objects in possible worlds that cannot be captured in the standard possible world semantics. The semantic analysis of (9) calls for a "cross-world" mode of evaluation.

In itself, the above point is independent of the question of whether $\bigcirc$ is extensional or not. However (9) may very well follow from an application of the principle of substitution *salva veritate.* Premises:

[3] Quine argues for this requirement in his [27]. Notoriously, Kripke [19] defended the view that (E-$\Box$) holds for constants (proper names are rigid designators). We do not make this assumption in this paper.

[4] Since then, Pluto is no longer considered a planet of the solar system (cf. `https://www.iau.org/public/themes/pluto`)

(10) Sam has an obligation not to be a tyrant: $\bigcirc\neg T(s)$

(11) Sam is a tyrant: $s = \iota x T(x)$

Conclusion:

(12) The tyrant has an obligation not to be a tyrant: $\bigcirc\neg T(\iota x T(x))$

One could object that (9) may alternatively be rendered as $\exists x(T(x) \wedge \bigcirc\neg T(x))$. This formalisation is unproblematic. First, we point out that as a spin-off of the extensionality of the deontic operator the principles of universal instantiation and existential generalisation hold unrestrictedly (viz. even if t is inside the scope of a deontic operator).

$$\exists x(x = t) \rightarrow (\forall x\varphi(x) \rightarrow \varphi(t)) \tag{UI}$$

$$\exists x(x = t) \rightarrow (\varphi(t) \rightarrow \exists x\varphi(x)) \tag{EG}$$

Given the assumption $\exists x(x = \iota y T(y))$, the two formalisations are equivalent. Thus the principle of extensionality turns an apparently unproblematic formula $(\exists x(T(x) \wedge \bigcirc\neg T(x))$ into a problematic one $(\bigcirc\neg T(\iota x T(x)))$. Our task is to account of the meaningfulness of the later formula. The following two derivations show the equivalence between the two formalisations. We use $\exists!$ for the uniqueness quantification defined as $\exists! x\varphi := \exists x\forall y(\varphi \leftrightarrow y = x)$.

$$\begin{array}{ll}
\text{(a) } \exists x(x = \iota y T(y)) & \text{(Hypothesis)} \\
\text{(b) } \exists x(T(x) \wedge \bigcirc\neg T(x)) & \text{(Hypothesis)} \\
\text{(c) } \exists! x T(x) & \text{(a)} \\
\text{(d) } \exists! x(T(x) \wedge \bigcirc\neg T(x)) & \text{(FO + b + c)} \\
\text{(e) } \forall x(T(x) \rightarrow \bigcirc\neg T(x)) & \text{(FO + d)} \\
\text{(f) } T(\iota y T(y)) \rightarrow \bigcirc\neg T(\iota y T(y)) & \text{(e + UI)} \\
\text{(g) } T(\iota y T(y)) & \text{(a)} \\
\text{(h) } \bigcirc\neg T(\iota y T(y)) & \text{(f + g)}
\end{array}$$

Derivation 1

$$\begin{array}{ll}
\text{(a) } \exists x(x = \iota y T(y)) & \text{(Hypothesis)} \\
\text{(b) } \bigcirc\neg T(\iota y T(y)) & \text{(Hypothesis)} \\
\text{(c) } T(\iota y T(y)) & \text{(a)} \\
\text{(d) } T(\iota y T(y)) \wedge \bigcirc\neg T(\iota y T(y)) & \text{(b + c)} \\
\text{(e) } \exists x(T(x) \wedge \bigcirc\neg T(x)) & \text{(d + EG)}
\end{array}$$

Derivation 2

2.2 Collapse

We explain in more detail how the collapse mentioned in requirement 3 arises. The discussion draws on Goble [12,13,14]. We say the deontic collapse arises in a logic if the formula $\varphi \leftrightarrow \bigcirc\varphi$ is derivable (for every formula φ). This would mean that everything that is true is obligatory and vice versa. Goble pointed out that, if the principle of substitution *salva veritate* holds, then the deontic collapse follows. We reiterate and amplify his main points.

The derivation of $\bigcirc\varphi \to \varphi$ presupposes that of $\varphi \to \bigcirc\varphi$. We start with the former one. It appeals to the law of contraposition, the law of double negation elimination, and the **D** axiom for $\bigcirc$:

(a) $\bigcirc\varphi$	(Hypothesis)
(b) $\neg\bigcirc\neg\varphi$	(**D** axiom)
(c) $\neg\neg\varphi$	($\varphi \to \bigcirc\varphi$ and contraposition)
(d) φ	(Double $\neg$ elimination)

Derivation 3

One may be tempted to block this derivation by just abandoning the principle of contraposition or the principle of double $\neg$ elimination. However, this would not block the derivation of $\varphi \to \bigcirc\varphi$, which in itself is counter-intuitive. We turn to this implication. We do not give the original argument, but a variant one, which highlights the role of $\Box$.

Proposition 2.1 *Consider a deontic logic containing (i) the usual principles of first-order logic (FO), (ii) the principle of substitution salva veritate for "ought" (E-$\bigcirc$), $t = s \to (\bigcirc\varphi \leftrightarrow \bigcirc\varphi_{t\hookrightarrow s})$ (iii) the principle $\Box\varphi \to \bigcirc\varphi$ ($\Box 2\bigcirc$) and (iv) the principle of inheritance "If $\vdash \varphi \to \psi$ then $\vdash \bigcirc\varphi \to \bigcirc\psi$" (In). Then $\varphi \to \bigcirc\varphi$ is derivable from $\Box\exists y(y = t)$.*

Proof. In this derivation we assume that x and y do not occur free in φ:

(a) φ	(Hypothesis)
(b) $\Box\exists y(y = t)$	(Hypothesis)
(c) $t = \imath x(x = t \land \varphi)$	(FO + a)
(d) $\bigcirc\exists y(y = t)$	($\Box 2\bigcirc$ + b)
(e) $\bigcirc\exists y(y = \imath x(x = t \land \varphi))$	(E-$\bigcirc$ + c + d)
(f) $\bigcirc\varphi$	(In + e)

Derivation 4

□

Some comments are in order:

- We show $\varphi \to \bigcirc\varphi$, where the original argument shows $\bigcirc\psi \to (\varphi \to \bigcirc\varphi)$.
- Our derivation starts from the supposition $\Box\exists y(y = t)$. This may be read as t necessarily denotes. We take this supposition to be harmless. We do not even want the collapse under this assumption.
- Line (c) "drags" φ inside the scope of the definite description to write "the-unique-x-identical-with-t-and-φ". Line (f) "drags" φ outside the scope of the definite description. The move is allowed in first-order logic.
- The principle (E-$\bigcirc$) is used on line (d), where t is replaced by the co-referential term "the-unique-x-identical-with-t-and-φ". The formula (e)

seems already counter-intuitive. However, as we will see in Section 4.3 the two-dimensional semantics presented in this paper gives an unproblematic reading to this formula.

- Line (f) is obtained by applying (In). This final move is explained in more detail in derivation 5.

To avoid the deontic collapse, the following ways out suggest themselves:

Option 1 Revise the laws of first-order logic;

Option 2 Abandon $(\Box 2\bigcirc)$;

Option 3 Abandon (In), or restrict its application.

We will go with option 3. Thus, in derivation 4, the move from (e) to (f) is blocked. A good reason for choosing this path is that option 2 alone would not block the original derivation of the collapse in a mono-modal setting, which uses (In) and the laws of first-order logic. Note that in Åqvist's system **F**, (In) is not a primitive rule, but is derivable from $(\Box 2\bigcirc)$ and two extra principles:

- the principle of necessitation for $\Box$: "If $\vdash \varphi$, then $\vdash \Box\varphi$" (N-$\Box$)
- the K axiom for $\bigcirc$: $\bigcirc(\varphi \rightarrow \psi) \rightarrow (\bigcirc\varphi \rightarrow \bigcirc\psi)$ (K-$\bigcirc$)

This is easily verified. The move from (e) to (f) is explained thus:

$$\begin{array}{ll}
\text{(a)} \vdash \exists y(y = \iota x(x = t \wedge \varphi)) \rightarrow \varphi & \text{(FO)} \\
\text{(b)} \vdash \Box[\exists y(y = \iota x(x = t \wedge \varphi)) \rightarrow \varphi] & \text{(N-}\Box\text{)} \\
\text{(c)} \vdash \bigcirc[\exists y(y = \iota x(x = t \wedge \varphi)) \rightarrow \varphi] & (\Box 2\bigcirc) \\
\text{(d)} \vdash \bigcirc\exists y(y = \iota x(x = t \wedge \varphi)) \rightarrow \bigcirc\varphi & \text{(K-}\bigcirc\text{)}
\end{array}$$

Derivation 5

Ultimately, the solution will consist in restricting the application of (N-$\Box$). However, the final effect will be the same: (In) will go away in its plain form. Prop. 2.2 provides an independent argument for keeping $\Box$ intensional (cf. requirement 2):

Proposition 2.2 *Consider the same deontic logic as in Prop. 2.1, but with (E-$\bigcirc$) replaced with (E-$\Box$). In such a logic, $\varphi \rightarrow \bigcirc\varphi$ is derivable from $\Box\exists y(y = t)$.*

Proof. As before we assume that x and y do not occur free in φ:

$$\begin{array}{ll}
\text{(a) } \varphi & \text{(Hypothesis)} \\
\text{(b) } \Box\exists y(y = t) & \text{(Hypothesis)} \\
\text{(c) } t = \iota x(x = t \wedge \varphi) & \text{(FO + a)} \\
\text{(d) } \Box\exists y(y = \iota x(x = t \wedge \varphi)) & \text{(E-}\Box\text{ + b + c)} \\
\text{(e) } \bigcirc\exists y(y = \iota x(x = t \wedge \varphi)) & (\Box 2\bigcirc) \\
\text{(f) } \bigcirc\varphi & \text{(In)}
\end{array}$$

Derivation 6

□

3 The perspectival account

In this section, we develop in full detail our perspectival account. The basic idea is that the content of an obligation at one world is to be evaluated from the perspective of another one. What we mean by this is the following. Formulas will be evaluated with respect to two dimensions, or pair of worlds (v, w). World v is where the evaluation takes place, and world w is the one from the perspective of which formulas are evaluated (call it the reference or actual world, if you wish). Throughout the paper the reference world will be represented as an upper index in the notation $v \models^w$. What is meant by "φ is evaluated in v from w's perspective" is this: when determining the truth-value of φ in v, the terms occurring in φ get the same denotation as in w.
To get a more flexible framework, we introduce two alethic modal operators, $\boxdot$ and $\boxtimes$. The first will be extensional, and the second intensional. Our prime interest is in combining extensionality for $\bigcirc$ and intensionality for $\Box$. However, there are readings of $\Box$ under which extensionality remains desirable. Hence we allow for both.

Definition 3.1 The language $\mathcal{L}$ contains:

- A countable set of variables $V := \{x, y, z, ...\}$
- A countable set of constants $C := \{c, d, e, ...\}$
- Two propositional connectives $\wedge, \neg$
- Three first-order logic symbols $\forall, \iota, =$
- A binary obligation operator $\bigcirc(-/-)$
- Two unary alethic operators $\boxdot$ and $\boxtimes$
- For each $n \in \mathbb{Z}^+$ a countable set of n-place predicate symbols $\mathbb{P} := \{A^n, B^n, ...\}$

We can now define inductively the well-formed terms and formulas used in our logic and their respective complexity ($\ulcorner ... \urcorner$).

Definition 3.2 [Terms and formulas]

- **Terms:**
  - · Every element of $V \cup C$ is a term of complexity 0
  - · If φ is a formula and $x \in V$ then $\iota x\varphi$ is a term with $\ulcorner \iota x\varphi \urcorner := \ulcorner \varphi \urcorner + 1$
- **Formulas:**
  - · If $R^n \in \mathbb{P}$ is a n-place predicate symbol and $t_1, ..., t_n$ are terms then $R^n(t_1, ..., t_n)$ is a formula with $\ulcorner R^n(t_1, ..., t_n) \urcorner := \sum_{i=1}^{n} \ulcorner t_i \urcorner$
  - · If φ is a formula and $x \in V$ then $\forall x\varphi$ is a formula with $\ulcorner \forall x\varphi \urcorner := \ulcorner \varphi \urcorner + 1$
  - · If t_1 and t_2 are terms then $t_1 = t_2$ is a formula with $\ulcorner t_1 = t_2 \urcorner := \ulcorner t_1 \urcorner + \ulcorner t_2 \urcorner + 1$
  - · If φ is a formula then $\neg\varphi$ is a formula with $\ulcorner \neg\varphi \urcorner := \ulcorner \varphi \urcorner + 1$
  - · If φ is a formula then $\boxdot\varphi$ is a formula with $\ulcorner \boxdot\varphi \urcorner := \ulcorner \varphi \urcorner + 1$
  - · If φ is a formula then $\boxtimes\varphi$ is a formula with $\ulcorner \boxtimes\varphi \urcorner := \ulcorner \varphi \urcorner + 1$
  - · If φ and ψ are formulas then $\varphi \wedge \psi$ is a formula with $\ulcorner \varphi \wedge \psi \urcorner := \ulcorner \varphi \urcorner + \ulcorner \psi \urcorner + 1$

· If φ and ψ are formulas then $\bigcirc(\psi/\varphi)$ is a formula
with $\ulcorner\bigcirc(\psi/\varphi)\urcorner := \ulcorner\varphi\urcorner + \ulcorner\psi\urcorner + 1$

· Nothing else is a formula

Definition 3.3 [Derived connectives] Let t be a term. We define $E(t)$ as $\exists x(x = t)$, where x is the first element of V not appearing in t. The symbols $\vee, \bot, \top, \rightarrow, \leftrightarrow, \Diamonddot\varphi, \diamondtimes\varphi, P(./.), \exists, \exists!$ and $\neq$ are introduced the usual way.

Definition 3.4 [Frames] $\mathcal{F} = \langle W, \succeq, D\rangle$ is called a frame, where

- $W \neq \emptyset$ is a set of worlds
- $\succeq \subseteq W \times W$ is a binary relation called the betterness relation [5]
- D is a function which maps every world $w \in W$ to a non-empty set D_w

D is called the domain function, and D_w is called the domain of w.
$\mathbb{D} := \bigcup_{w \in W} D_w$ is called the "actual" domain and $\mathbb{D}^+ := \mathbb{D} \cup \{\mathbb{D}\}$ the (whole) domain.

The individual domains $(D_w)_{w \in W}$ contain all objects which are within the range of the universal quantifier at a world w. The actual domain $\mathbb{D}$ is not contained in the domain of any world [6] and is used as the value assigned to definite descriptions that do not designate (uniquely).

Definition 3.5 [Models] $\mathcal{M} = \langle W, \succeq, D, I\rangle$ is called a model (on the frame $\mathcal{F} = \langle W, \succeq, D\rangle$), where I is a function (called interpretation function) such that:

- for $c \in C$ and $w \in W$: $I(c, w) \in \mathbb{D}^+$
- for $R^n \in \mathbb{P}$ and $w \in W$: $I(R^n, w) \subseteq (\mathbb{D}^+)^n$

$I(c, w) = a$ says that a is the denotation of c in w. In our semantics a constant may not denote, and it does not need to designate the same entity in every possible world. In Kripke's terminology, proper names are not rigid designators. We have not investigated the effects of making this assumption.

Definition 3.6 [Variable assignment] Given a model $\mathcal{M} = \langle W, \succeq, D, I\rangle$ we call a function $g : V \times W \mapsto \mathbb{D}^+$ a variable assignment (of $\mathcal{M}$).

Notice that the assignment is world-dependent. Roughly speaking, $g(x, w) = a$ says that a is the denotation of x in w. Note that $g(x, w)$ does not have to be an element of the domain of w. [7] We amend the usual notion of an x-variant as follows. An *x-variant of some variable assignment g at a world w* is a variable assignment h that agrees with g on all values except for x, whose value in every world remains constant, and an element of D_w. Formally:

Definition 3.7 [x-variant] Assume a model $\mathcal{M} = \langle W, \succeq, D, I\rangle$, a variable assignment g of $\mathcal{M}$ and an element of the whole domain $d \in \mathbb{D}^+$. We write $g_{x \Rightarrow d}$

[5] When $w \succeq v$, we say that a world w is at least as good as world v.

[6] $\mathbb{D} \notin \mathbb{D}$.

[7] The element a does not even have to be contained in the actual domain.

for the variable assignment which replaces the value assigned to x at any world by d:

$$g_{x\Rightarrow d}(z,v) := \begin{cases} d & \text{if } \ (z,v) \in \{x\} \times W \\ g(z,v) & \text{otherwise} \end{cases}$$

A variable assignment h is an *x-variant of g at w* iff $h = g_{x\Rightarrow d}$ for some $d \in D_w$.

"Best", in terms of which the truth-conditions for $\bigcirc(-/-)$ are cast, is defined by:

Definition 3.8 [best] Given a model $\mathcal{M} = \langle W, \succeq, D, I\rangle$ and a set of worlds $X \subseteq W$ we define

$$\mathit{best}(X) := \{w \in X : \forall v \in W(v \in X \Rightarrow w \succeq v)\}$$

$\mathit{best}(X)$ is the set of worlds in X that are at least as good as every member of X.

The construct "$\mathcal{M}, v \models^w_g \varphi$" can be read as "$v$ forces φ under g if looked at from the point of view of (an inhabitant of) w". We stress that $\mathcal{M}, v \models^w_g$ does not convey a truth value for the formula φ per se, but it is used to define the truth conditions of φ by induction. We put $||\varphi||^{\mathcal{M}}_{g,w} := \{v \in W : \mathcal{M}, v \models^w_g \varphi\}$.

Definition 3.9 Let $\mathcal{M} = \langle W, \succeq, D, I\rangle$ be a model, g a variable assignment, $x \in V$ and $c \in C$. We define

- $I^w_g(x) := g(x,w)$
- $I^w_g(c) := I(c,w)$
- $I^w_g(\iota x\varphi) := \begin{cases} h(x,w) & \text{if } h \text{ is the } \textbf{unique } x\text{-variant of } g \text{ at } w \\ & \text{such that } \mathcal{M}, w \models^w_h \varphi \\ \mathbb{D} & \text{otherwise} \end{cases}$

The forcing relation $\models$ can be defined inductively as follows:

- $\mathcal{M}, v \models^w_g R^n(t_1, ..., t_n) :\Leftrightarrow \langle I^w_g(t_1), ..., I^w_g(t_n)\rangle \in I(R^n, v)$
- $\mathcal{M}, v \models^w_g \neg\varphi :\Leftrightarrow \mathcal{M}, v \not\models^w_g \varphi$
- $\mathcal{M}, v \models^w_g \varphi \wedge \psi :\Leftrightarrow \mathcal{M}, v \models^w_g \varphi \ \text{ and } \ \mathcal{M}, v \models^w_g \psi$
- $\mathcal{M}, v \models^w_g \forall x\varphi :\Leftrightarrow \mathcal{M}, v \models^w_h \varphi \quad \text{for all } x\text{-variants } h \text{ of } g \text{ at } v$
- $\mathcal{M}, v \models^w_g t_1 = t_2 :\Leftrightarrow I^w_g(t_1) = I^w_g(t_2)$
- $\mathcal{M}, v \models^w_g \boxdot\varphi :\Leftrightarrow \forall u \in W \ \mathcal{M}, u \models^w_g \varphi$
- $\mathcal{M}, v \models^w_g \boxtimes\varphi :\Leftrightarrow \forall u \ \forall v' \in W \ \mathcal{M}, u \models^{v'}_g \varphi$
- $\mathcal{M}, v \models^w_g \bigcirc(\psi/\varphi) :\Leftrightarrow \mathit{best}(||\varphi||^{\mathcal{M}}_{g,w}) \subseteq ||\psi||^{\mathcal{M}}_{g,w}$

We drop the reference to $\mathcal{M}$ when it is clear what model is intended.

Definition 3.10 [Truth in $\mathbf{F}^{\forall}$] Given a model $\mathcal{M} = \langle W, \succeq, D, I\rangle$, a variable assignment g, a formula φ and a world w we define what it means that φ is true in $\mathcal{M}$ at w under g (in symbols: $\mathcal{M}, w \models_g \varphi$) as

$$\mathcal{M}, w \models_g \varphi :\Leftrightarrow \mathcal{M}, w \models^w_g \varphi$$

The meaning of $\boxdot$, $\boxtimes$ and $\bigcirc$ is easier to explain using the following derived truth conditions.

Remark 3.11 [Derived truth conditions]

- $\mathcal{M}, w \models_g \boxdot\varphi :\Leftrightarrow \forall v \in W\ \mathcal{M}, v \models_g^w \varphi$
- $\mathcal{M}, w \models_g \boxtimes\varphi :\Leftrightarrow \forall u\ \forall v \in W\ \mathcal{M}, u \models_g^v \varphi$
- $\mathcal{M}, w \models_g \bigcirc(\psi/\varphi) :\Leftrightarrow best(||\varphi||_{g,w}^{\mathcal{M}}) \subseteq ||\psi||_{g,w}^{\mathcal{M}}$

When evaluating the truth-value of $\boxdot\varphi$ at w, one moves to an arbitrary world v, and determines the truth-value of φ in v from w's perspective. This means giving to the terms occurring in φ the denotation they have in w. When evaluating the truth-value of $\boxtimes\varphi$ at w, one moves to an arbitrary world u, and evaluates φ in u from every other world's v perspective. For obligation, the idea is similar. The standard evaluation rule puts $\bigcirc(\psi/\varphi)$ as true whenever all the best φ-worlds are ψ-worlds. The φ-worlds and the ψ-worlds in question are those according to w's perspective. This is how the principle of substitution *salva veritate* will be validated for $\bigcirc$ and $\boxdot$, and invalidated for $\boxtimes$.

Definition 3.12 Given a model $\mathcal{M} = \langle W, \succeq, D, I \rangle$. $\succeq$ is reflexive if $\forall w \in W (w \succeq w)$, and $\succeq$ fulfils the limitedness condition if for every φ, g and $w \in W$ we have

$$||\varphi||_{g,w}^{\mathcal{M}} \neq \emptyset \Rightarrow best(||\varphi||_{g,w}^{\mathcal{M}}) \neq \emptyset$$

$\mathcal{U}$ is the class of models in which $\succeq$ is reflexive and fulfils limitedness.

Intuitively, the limitedness condition validates the dyadic version of the **D** axiom (with $\Diamond$ replaced with $\Diamonddot$) involved in derivation 3 of the collapse (see Subsect. 2.2).

Definition 3.13 [Validity in $\mathbf{F}^{\forall}$]

- φ is valid at w in a model $\mathcal{M}$ (notation: $\mathcal{M}, w \models \varphi$) if for every variable assignment g, we have that $\mathcal{M}, w \models_g \varphi$;
- φ is valid in a model $\mathcal{M}$ (notation: $\mathcal{M} \models \varphi$) if for every world w we have $\mathcal{M}, w \models \varphi$;
- φ is valid in a class $\mathbb{M}$ of models (notation: $\mathbb{M} \models \varphi$) if for every model $\mathcal{M} \in \mathbb{M}$ we have $\mathcal{M} \models \varphi$;
- φ is valid (notation: $\models \varphi$) if φ is valid in the class $\mathcal{U}$ as defined above.

4 Benchmarking

We test the account introduced in Sect. 3 against the requirements discussed in Sect. 2.

4.1 Extensionality / intensionality / self-negation

A proof of the principle of extensionality in its general form is given in Subsect. 4.2. For simplicity's sake, here we only discuss the examples considered in Sect. 2.

Proposition 4.1 (Extensionality of $\bigcirc$, requirement 1) *We have:*

$j = \imath x P(x) \rightarrow (\bigcirc(S(\imath x P(x)) \leftrightarrow \bigcirc S(j))$

$j = \imath xP(x) \rightarrow [\bigcirc(El(\imath y(y \neq \imath xP(x)))/\neg S(\imath xP(x))) \leftrightarrow \bigcirc(El(\imath y(y \neq \imath xP(x)))/\neg S(j))]$

Proof. When a formula does not contain a free variable its truth condition does not depend on which variable assignment is assumed. Therefore for this and all future proofs (in which no free variable is involved) we always deal with an arbitrary variable assignment. Now, if $w \models^w_g j = \imath xP(x)$, then for every $u \in best(||\top||^{\mathcal{M}}_{g,w})$

$$u \models^w_g S(\imath xP(x)) \Leftrightarrow u \models^w_g S(j)$$

This is because the terms on both sides get the denotation they have in w. Similarly:

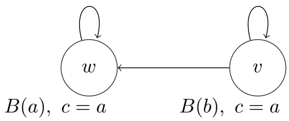
$$\begin{aligned} &best(||\neg S(\imath xP(x))||^{\mathcal{M}}_{g,w}) \subseteq ||El(\imath y(y \neq \imath xP(x)))||^{\mathcal{M}}_{g,w} \\ \Leftrightarrow\ &best(||\neg S(j)||^{\mathcal{M}}_{g,w}) \subseteq ||El(\imath y(y \neq \imath xP(x)))||^{\mathcal{M}}_{g,w} \end{aligned}$$

□

Proposition 4.2 (Intensionality of $\boxtimes$, requirement 2) *We do not have:*

$$c = \imath xB(x) \rightarrow (\boxtimes(c = c) \leftrightarrow \boxtimes(c = \imath xB(x)))$$

Proof. Put $\mathcal{M} = \langle W, \succeq, I, D\rangle$ with (an arrow from v to w means $v \succeq w$, and no arrow from from w to v means $w \not\succeq v$):



$W := \{w, v\}$
$\succeq :=$ the reflexive closure of $\{(v, w)\})$
$D_w := \{a\},\quad D_v := \{a, b\}$
$I(B, w) := a,\quad I(B, w) := b$
$I(c, w) := a,\quad I(c, v) := a$

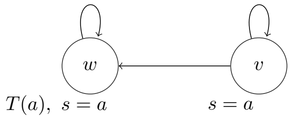
The condition of limitedness is fulfilled. We have:

- $w \models^w_g c = \imath xB(x)$ since c and $\imath xB(x)$ denote a in w
- $w \models^w_g \boxtimes(c = c)$ since $c = c$ is a tautology
- $w \not\models^w_g \boxtimes(c = \imath xB(x))$ since $w \not\models^v_g c = \imath xB(x)$[8]

□

Proposition 4.3 (Self-negation, requirement 4) *The sentences (10), (11) and (12) are simultaneously satisfiable.*

Proof. We give a model which satisfies all three formulas in the same world.



$W := \{w, v\}$
$\succeq :=$ the reflexive closure of $\{(v, w)\})$
$D_w := \{a\},\quad D_v := \{a\}$
$I(T, w) := \{a\},\quad I(T, v) := \emptyset$
$I(s, w) := a,\quad I(s, v) := a$

[8] c and $\imath xB(x)$ do not have the same denotation in v.

As before $\succeq$ is limited. We have:

- $w \models_g^w s = \imath x T(x)$ since s and $\imath x T(x)$ denote a in w
- $w \models_g^w \bigcirc\neg T(s)$ since a is not T in v
- $w \models_g^w \bigcirc\neg T(\imath x T(x))$ since a (=the unique T in w) is not T in v

The paradox is resolved by having Sam, who is the tyrant in the actual world w, not be a tyrant in the best world v. Therefore $\bigcirc\neg T(\imath x T(x))$ can be satisfied. □

4.2 Extensionality (general form)

We show the principle of extensionality in its general form. Where φ is a formula and s and t terms, let $\varphi_{t\hookrightarrow s}$ be the result of replacing zero up to all unbound occurrences of t,[9] in φ, by s. We may re-letter bound variables, if necessary, to avoid rendering the new occurrences of variables in s bound in φ.

Proposition 4.4 *Consider some g and some w in $\mathcal{M}$ such that $w \models_g^w t = s$. Then, for all v in $\mathcal{M}$,*

$$v \models_g^w \varphi \leftrightarrow \varphi_{t\hookrightarrow s} \qquad (\#)$$

provided t is not contained in the scope of the $\boxtimes$ operator in φ.

Proof. By induction on the complexity n of a formula φ. The base case, if φ is $R(t_1, ..., t_m)$ with $\ulcorner R(t_1, ..., t_m)\urcorner = 0$, follows from the definitions involved. For the inductive case, we assume (#) holds for all $k < n$, and *for all* v in $\mathcal{M}$. We only consider three cases−the other ones are left to the reader:

- $\varphi := \forall x\ \psi$. Given the restrictions put on t and s, we have the following chain of equivalences:

$$\begin{aligned} v \models_g^w \forall x\ \psi \text{ iff } & v \models_h^w \psi \quad \text{for all } x\text{-variants } h \text{ at } v \\ & v \models_h^w \psi_{t\hookrightarrow s} \quad \text{for all } x\text{-variants } h \text{ at } v \text{ (by IH)} \\ & v \models_g^w \forall x\ \psi_{t\hookrightarrow s} \end{aligned}$$

- $\varphi := \bigcirc(\chi/\psi)$.

$$\begin{aligned} v \models_g^w \bigcirc(\chi/\psi) \text{ iff } & best(||\psi||_{g,w}^{\mathcal{M}}) \subseteq ||\chi||_{g,w}^{\mathcal{M}} \\ & best(||\psi_{t\hookrightarrow s}||_{g,w}^{\mathcal{M}}) \subseteq ||\chi_{t\hookrightarrow s}||_{g,w}^{\mathcal{M}} \text{ (by IH)} \\ & v \models_g^w \bigcirc(\chi_{t\hookrightarrow s}/\psi_{t\hookrightarrow s}) \\ & v \models_g^w \bigcirc(\chi/\psi)_{t\hookrightarrow s} \end{aligned}$$

- $\varphi := R(t_1, ..., t_m)$. Assume $v \models_g^w R(t_1, ..., t_m)$. If t appears only as one of the t_i's, then we are done. So let us suppose that t appears *in* one (or more) of the t_i's. W.l.o.g. let t only appear in $t_1 = \imath x\psi$. By the IH $w \models_g^w \psi \leftrightarrow \psi_{t\hookrightarrow s}$, so $I_g^w(\imath x\psi) = I_g^w(\imath x\psi_{t\hookrightarrow s})$. Consider some $v \in W$.

[9] By an unbounded occurrence of t, we mean that no variables in t are in the scope of a quantifier or a definite description *not* in t.

We have $\langle I^w_g(\iota x\psi), ..., t_m\rangle \in I(R, v)$, so $\langle I^w_g(\iota x\psi_{t\hookrightarrow s}), ..., t_m\rangle \in I(R, v)$. Hence $v \models^w_g R(t_1, ..., t_m)_{t\hookrightarrow s}$ as required. For the converse implication, the argument is the same.

□

Corollary 4.5 (Extensionality) *The principle (E) is valid:*

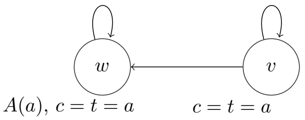
$$\models t = s \to (\varphi \leftrightarrow \varphi_{t\hookrightarrow s}) \quad \textit{if } t \textit{ is not in the scope of } \boxtimes \tag{E}$$

Proof. This follows from Prop. 4.4 putting $v = w$. □

4.3 Deontic collapse

We start by explaining how the collapse is avoided semantically. We define a model in which the formulas at steps (a)-(e) in derivation 4 are true in the actual world w but the formula at step (f) is not.

Example 4.6 Put $\varphi := A(c)$. $\mathcal{M}$ is defined by



$W := \{w, v\}$
$\succeq :=$ the reflexive closure of $\{(v, w)\}$
$D_w := \{a\}$, $D_v := \{a\}$
$I(c, w) := I(c, v) := a$
$I(t, w) := I(t, v) := a$
$I(A, w) := \{a\}$, $I(A, v) := \emptyset$

We have

(a) $w \models_g A(c)$ since $I(c, w) = a \in I(A, w)$
(b) $w \models_g \boxtimes\exists y(y = t)$ since $I(t, w) = I(t, v) = a \in D_w$
and $I(t, w) = I(t, v) = a \in D_v$
(c) $w \models_g t = \iota x(x = t \wedge A(c))$ since $I(t, w) = a = I^w_g(\iota x(x = t \wedge A(c)))$
(d) $w \models_g \bigcirc\exists y(y = t)$ since $I(t, w) = a \in D_v$[10]
(e) $w \models_g \bigcirc\exists y(y = \iota x(x = t \wedge A(c)))$ since $I^w_g(\iota x(x = t \wedge A(c))) = a \in D_v$
(f) $w \not\models_g \bigcirc A(c)$ since $I(c, w) = a \notin I(A, v)$

Let it be clear that (e) means $v \models^w_g \exists y(y = \iota x(x = t \wedge A(c)))$, which says that the unique x, for which the formula $x = t \wedge A(c)$ holds in w, exists in v. However this does NOT imply $v \models^w_g \exists x(x = t \wedge A(c)))$, since there exists no element in the domain of v for which the formula $x = t \wedge A(c)$ holds in v from w's perspective. In the statements, $v \models^w_g \exists y(y = \iota x(x = t \wedge A(c)))$ and $v \models^w_g \exists x(x = t \wedge A(c))$ the two c refer to the same individual a, but in different worlds where they have different properties.

This model serves as a counter-model to the rule of inheritance. The formula $\exists y(y = \iota x(x = t \wedge A(c))) \to A(c)$ is valid, but not $\bigcirc\exists y(y = \iota x(x = t \wedge A(c))) \to \bigcirc A(c)$.

[10] By definition $v \models^w_g \exists y(y = t)$ holds if there exists an y-variant h of g at v such that $h(y, w) = I(t, w)$. This is equivalent to $I(t, w)$ being an element of D_v.

To explain how the deontic collapse is avoided proof-theoretically, we introduce the notion of "variable only" version φ^* of a formula φ. Intuitively, φ^* is obtained by substituting, in φ, a new variable for every definite description and constant occurring in φ. This ensures that φ^* contains only variables, making it impossible to apply the rule of inheritance (and necessitation) from which the collapse follows. Formally:

Definition 4.7 [Variable only version, Goble [13]] Given a formula φ, we define φ^* as the formula in which all terms $t_1, ..., t_n$, which are not variables and are occurring in the formula φ, have been replaced by $x_1, ..., x_n \in V$ respectively. The variables $x_1, ..., x_n$ are the first, pairwise different, elements of V such that $x_1, ..., x_n$ do not occur in φ.

Example 4.8 Let A, B and C be predicate symbols, $x, y, z \in V$ the first three variables of V, $c \in C$ a constant and $\varphi \in WF$ a well-formed formula:

- $A(\iota y\varphi, c)^* = A(x, z)$
- $\forall x A(\iota y B(y, d), x)^* = \forall x A(z, x)$
- $A(\iota y B(\iota x C(x, y)), y)^* = A(z, y)$
- $A(y, y)^* = A(y, y)$

Like in Goble's original treatment, the collapse is blocked by restricting the application of the rule of necessitation for $\boxtimes$, and of the principle of inheritance for $\bigcirc$. These two are now available in the following form:

$$\text{If } \models \varphi^* \text{ then } \models \boxtimes\varphi \qquad (\text{N}^*\text{-}\boxtimes)$$
$$\text{If } \models (\psi_1 \to \psi_2)^* \text{ then } \models \bigcirc(\psi_1/\varphi) \to \bigcirc(\psi_2/\varphi)) \qquad (\text{In}^\star)$$

Before continuing want to point out that the other law involved in the collapse, $\boxtimes\psi \to \bigcirc(\psi/\varphi)$, still holds. This follows at once from the following:

Proposition 4.9 *We have*

$$\models \boxtimes\psi \to \boxdot\psi \qquad (\boxtimes 2\boxdot)$$
$$\models \boxdot\psi \to \bigcirc(\psi/\varphi) \qquad (\boxdot 2\bigcirc)$$

Proof. $(\boxtimes 2\boxdot)$ is straightforward, and may be left to the reader. For $(\boxdot 2\bigcirc)$, let us assume $w \models_g \boxdot\psi$ holds for a fixed model $\mathcal{M} = \langle W, \succeq, D, I\rangle$, a world $w \in W$ and a variable assignment g. This is equivalent to $||\psi||^{\mathcal{M}}_{g,w}$ being equal to the whole set of worlds W. Hence we can infer that for any formula φ we have $best(||\varphi||^{\mathcal{M}}_{g,w}) \subseteq W = ||\psi||^{\mathcal{M}}_{g,w}$, which, by definition, means $w \models_g \bigcirc(\psi/\varphi)$. □

We now show that the rules $(\text{N}^*\text{-}\boxtimes)$ and $(\text{In}^\star)$ preserve validity. To show this we need the following two lemmas.

Lemma 4.10 *Given a formula φ and a model $\mathcal{M}$, then*

$$\mathcal{M} \models \varphi^* \Rightarrow \mathcal{M} \models \boxtimes(\varphi^*)$$

Proof. Let φ be a formula and $\mathcal{M} = \langle W, \succeq, D, I\rangle$ a model. If for every world $w \in W$ and every variable assignment g of $\mathcal{M}$ it holds that $w \models_g \varphi^*$, it follows that $w \models^w_g \varphi^*$ holds for every world $w \in W$ and every variable assignment g of $\mathcal{M}$. Now let us take two arbitrary but fixed worlds $v, w \in W$ and an arbitrary but fixed variable assignment g and define a new variable assignment $h : V \times W \to \mathbb{D}^+$ of $\mathcal{M}$ as:

$$h(x,u) := \begin{cases} g(x,w) & \text{if } \ u = v \\ g(x,v) & \text{if } \ u = w \\ g(x,u) & \text{otherwise} \end{cases}$$

Since h and g only swap how they see the variables at w and v, and φ^* does not contain constants or definite descriptions, we get $\forall u(u \models_g^w \varphi^* \Leftrightarrow u \models_h^v \varphi^*)$. Therefore from $v \models_h^v \varphi^*$, which holds by assumption, we can infer $v \models_g^w \varphi^*$. Since $v, w \in W$ and g were arbitrary we can conclude $\mathcal{M} \models \boxtimes\varphi^*$. □

Lemma 4.11 *Given a formula φ and a model $\mathcal{M}$, then*

$$\mathcal{M} \models \varphi^* \Rightarrow \mathcal{M} \models \varphi$$

Proof. This proof is done by contraposition. Suppose there are $\mathcal{M} = \langle W, \succeq, D, I \rangle$, $w \in W$ and g such that $w \not\models_g^w \varphi$. Let $t_1, ..., t_n$ be all terms in φ which are replaced by the corresponding variables $x_1, ..., x_n$ in φ^*. Then for the variable assignment

$$h(x,v) := \begin{cases} I_g^v(t_i) & \text{if } \ (x,v) \in \{x_i\} \times W \ \ \text{where } \ i \in \{1,...,n\} \\ g(x,v) & \text{otherwise} \end{cases}$$

we have $w \not\models_h^w \varphi^*$. □

Putting those two lemmas together, we can prove the soundness of (N*-$\boxtimes$):

Lemma 4.12 *Given a formula φ and a model $\mathcal{M}$ then*

$$\mathcal{M} \models \varphi^* \quad \textit{implies} \quad \mathcal{M} \models \boxtimes\varphi$$

Proof. $\mathcal{M} \models \varphi^* \Rightarrow \mathcal{M} \models \boxtimes(\varphi^*) \Leftrightarrow \mathcal{M} \models (\boxtimes\varphi)^* \Rightarrow \mathcal{M} \models \boxtimes\varphi$. □

Theorem 4.13 *We have*

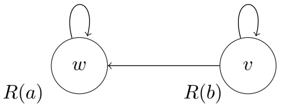
$$\textit{If } \models \varphi^* \textit{ then } \models \boxtimes\varphi \qquad (\text{N}^*\text{-}\boxtimes)$$

$$\textit{If } \models (\psi_1 \to \psi_2)^* \textit{ then } \models \bigcirc(\psi_1/\varphi) \to \bigcirc(\psi_2/\varphi) \qquad (\text{In}^\star)$$

Proof. The first rule follows at once from Lem. 4.12. The second rule follows from the first one and Prop. 4.9. □

We end with the observation that the rule of necessitation in its plain form fails for $\boxtimes$. Here is a counter-example. The formula $\exists y(y = \imath x R(x)) \to R(\imath x R(x))$ is valid in any model. To see why, fix a model $\mathcal{M} = \langle W, \succeq, D, I \rangle$, a variable assignment g, and a world $w \in W$. Assume $w \models_g \exists y(y = \imath x R(x))$. Hence, there exists a y-variant h of g at w such that $h(y,w) = I_h^w(\imath x R(x))$. This means that $h(y,w) = a$ for some $a \in D_w$. By definition of $\imath x R(x)$, a is the unique element in D_w s.t. $a \in I(R,w)$. So $w \models_h R(\imath x R(x))$. Since y does not occur in $R(\imath x R(x))$ we conclude $w \models_g R(\imath x R(x))$ as required.
Now we define a model in which $\boxtimes[\exists y(y = \imath x R(x)) \to R(\imath x R(x))]$ is not valid:

Example 4.14 Consider the model $\mathcal{M} := \langle W, \succeq, D, I \rangle$ with



$W := \{w, v\}$
$\succeq :=$ the reflexive closure of $\{(v, w)\}$
$D_w := \{a, b\}, \quad D_v := \{a, b\}$
$I(R, w) := \{a\}, \quad I(R, v) := \{b\}$

We have $v \models_g^w \exists y(y = \imath x R(x))$, as $I_g^w(\imath x R(x)) = a \in D_v$. But $v \not\models_g^w R(\imath x R(x))$ because $I_g^w(\imath x R(x)) = a \notin I(R, v)$. So $\mathcal{M} \not\models \boxtimes[\exists y(y = \imath x R(x)) \to R(\imath x R(x))]$.

5 Concluding remarks

We have defined and studied a new perspectival account of conditional obligation. A number of requirements were identified, and shown to be met by the framework. The framework allows for a more nuanced way of approaching first-order deontic principles.
Topics for future research include:

(i) to investigate variant candidate truth-conditions for $\boxtimes$

(ii) to find a suitable axiomatic basis

Ad (i): the truth-conditions for $\boxtimes$ in Def. 3.9 allowed us to make the minimal changes to the axiomatic basis of **F**. The most significant change is that Lewis's absoluteness principle $\bigcirc(\psi/\varphi) \to \boxtimes \bigcirc (\psi/\varphi)$, stipulating that obligations are necessary, goes away. This may be considered good news. But ($\boxtimes 2 \bigcirc$) remains, and this law may be considered counter-intuitive. The following alternative truth-conditions may be used:

$$w \models_g \boxtimes\varphi \text{ iff } \forall v: \ v \models_g^v \varphi$$

Intuitively: $w \models_g \boxtimes\varphi$ holds, if φ holds at all v under the hypothesis that the terms occurring in φ take the reference they have in this very same world. With this definition of $\boxtimes$, ($\boxtimes 2 \bigcirc$) goes away, and the rule of necessitation holds without any restriction.
Ad (ii): we have identified a sound axiomatic basis for the logic. This one is shown in Appendix B. Completeness is left as a topic for future research.

Appendix A: Åqvist's system F

Axioms:

All truth-functional tautologies

S5-schemata for $\Box$ and $\Diamond$

$\bigcirc\,(\varphi \to \chi/\psi) \to (\bigcirc(\varphi/\psi) \to \bigcirc(\chi/\psi))$

$\bigcirc\,(\varphi/\psi) \to \Box\bigcirc(\varphi/\psi)$

$\Box\varphi \to \bigcirc(\varphi/\psi)$

$\Box(\varphi \leftrightarrow \psi) \to (\bigcirc(\chi/\varphi) \leftrightarrow \bigcirc(\chi/\psi))$

$\bigcirc\,(\varphi/\varphi)$

$\bigcirc\,(\varphi/\psi \wedge \chi) \to \bigcirc(\chi \to \varphi/\psi)$

$\Diamond\psi \to (\bigcirc(\varphi/\psi) \to P(\varphi/\psi))$

Rules:

If $\vdash \varphi$ and $\vdash \varphi \to \chi$ then $\vdash \chi$

If $\vdash \varphi$ then $\vdash \Box\varphi$

An explanation of the axioms can be found in [24]. The dyadic version of the **D** axiom ($\Diamond\psi \to (\bigcirc(\varphi/\psi) \to P(\varphi/\psi))$) is the distinguishing axiom of this logic. This axiom makes the system **F** the weakest system in the family of Åqvist's systems in which the collapse arises.

Appendix B: Axiomatisation of $\mathbf{F}^{\forall}$

A sound Hilbert axiomatisation of the logic developed in this paper is shown below. In this axiomatisation, the symbol $\varphi_{x \Rightarrow t}$ is the result of replacing ALL occurrences of the variable x, in φ, by the term t. Furthermore, we write $free(\varphi)$ for the set of variables appearing in φ, which are not bound by a quantifier or a definite description.

Axioms:

All truth functional tautologies

All axioms of system **F** with $\Box$ replaced with $\boxdot$ and $\Diamond$ with ⟐

S5-schemata for $\boxtimes$ and $\diamondtimes$

$\boxtimes\,\varphi \to \boxdot\varphi$

$\boxtimes\,\psi \to \boxtimes\bigcirc(\psi/\varphi)$

$\boxtimes\,(\varphi \leftrightarrow \psi) \to \boxtimes(\bigcirc(\chi/\varphi) \leftrightarrow \bigcirc(\chi/\psi))$

$t = s \to (\varphi \leftrightarrow \varphi_{t \hookrightarrow s})$ if t is not in the scope of the $\boxtimes$ operator

$E(t) \to (\forall x\varphi \to \varphi_{x \Rightarrow t})$ if x is not in the scope of the $\boxtimes$ operator

$\exists x \exists y(x = y)$

$t = t$

$t \neq s \to \boxdot t \neq s$
$\forall y((\forall x(\varphi \leftrightarrow x = y)) \to y = \text{℩} x\varphi)$
$E(\text{℩} x\varphi) \to \exists! x\varphi$
$\forall x(E(x) \to \varphi) \to \forall x\varphi$
$(\forall x\varphi \wedge \forall x\psi) \leftrightarrow \forall x(\varphi \wedge \psi)$

Rules:
If $\vdash \varphi$ and $\vdash \varphi \to \chi$ then $\vdash \chi$
If $\vdash \varphi^*$ then $\vdash \boxtimes\varphi$
If $\vdash \bigcirc(\varphi/\psi)$ then $\vdash \boxtimes \bigcirc (\varphi/\psi)$
If $\vdash \varphi \to t \neq x$ then $\vdash \neg\varphi$ where $x \notin$ free(φ)
If $\vdash \varphi \to \psi$ then $\vdash \varphi \to \forall x\psi$ where $x \notin$ free(φ)
If $\vdash \varphi \to \boxdot\psi$ then $\vdash \varphi \to \boxdot\forall x\psi$ where $x \notin$ free(φ)
If $\vdash \varphi \to \boxtimes\psi$ then $\vdash \varphi \to \boxtimes\forall x\psi$ where $x \notin$ free(φ)

An explanation of the first-order and definite description axioms can be found in [30].

References

[1] Åqvist, L., "An Introduction to Deontic logic and the Theory of Normative Systems," Bibliopolis, Naples, 1987.

[2] Castañeda, H., *The paradoxes of deontic logic: The simplest solution to all of them in one fell swoop*, in: R. Hilpinen, editor, *New Studies in Deontic Logic*, Springer, 1981 pp. 37–85.

[3] Chalmers, D., *Epistemic two-dimensional semantics*, Philosophical Studies **118** (2004), p. 153–226.

[4] Chisholm, R. M., *Contrary-to-duty imperatives and deontic logic*, Analysis **24** (1963), pp. 33–36.

[5] Danielsson, S., "Preference and Obligation," Filosofiska Fäeningen, Uppsala, 1968.

[6] de Boer, M., D. M. Gabbay, X. Parent and M. Slavkovic, *Two dimensional standard deontic logic*, Synth. **187** (2012), pp. 623–660.

[7] Frijters, S., "All Doctors Have an Obligation to Care for Their Patients: Term-Modal Logics for Ethical Reasoning with Quantified Deontic Statements," Ph.D. thesis, Ghent (2021).
URL `http://hdl.handle.net/1854/LU-8698101`

[8] Frijters, S. and T. D. Coninck, *The Manchester twins: Conflicts between directed obligations*, in: F. Liu, A. Marra, P. Portner and F. V. D. Putte, editors, *Deontic Logic and Normative Systems - 15th International Conference, DEON 2020/21* (2021), pp. 166–182.

[9] Fusco, M., *A two-dimensional logic for diagonalization and the a priori*, Synth. **198** (2021), pp. 8307–8322.

[10] Fusco, M. and A. W. Kocurek, *A two-dimensional logic for two paradoxes of deontic modality*, Rev. Symb. Log. **15** (2022), pp. 991–1022.

[11] Gabbay, D. M., A. Kurucz, F. Wolter and M. Zakharyaschev, "Multi-dimensional modal logic," Elsevier, 2003.

[12] Goble, L., *Opacity and the ought-to-be*, Noûs (1973), pp. 407–412.

[13] Goble, L., *Quantified deontic logic with definite descriptions*, Logique et Analyse **37** (1994), pp. 239–253.

[14] Goble, L., *'Ought' and extensionality*, Noûs **30** (1996), pp. 330–355.

[15] Goble, L., *Axioms for Hansson's dyadic deontic logics*, Filosofiska Notiser **6** (2019), pp. 13–61.

[16] Hansson, B., *An analysis of some deontic logics*, Noûs (1969), pp. 373–398.

[17] Horty, J. F., *Perspectival act utilitarianism*, in: P. Girard, O. Roy and M. Marion, editors, *Dynamic Formal Epistemology*, Springer Netherlands, Dordrecht, 2011 pp. 197–221.

[18] Humberstone, L., *Two-dimensional adventures*, Philosophical Studies **118** (2004), pp. 17–65.

[19] Kripke, S., "Naming and Necessity," Harvard University Press, Cambridge, 1980.

[20] Kölbel, M., "Truth without Objectivity," Routledge, 2002.

[21] Lewis, D., "Counterfactuals," Blackwell, Oxford, 1973.

[22] MacFarlane, J., "Assessment Sensitivity: Relative Truth and its Applications," Oxford University Press, 2014.

[23] Parent, X., *Completeness of Åqvist's systems E and F*, The Review of Symbolic Logic **8** (2015), pp. 164–177.

[24] Parent, X., *Preference-based semantics for Hansson-type dyadic deontic logic*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *The Handbook of Deontic Logic and Normative Systems*, College Publications, 2021 pp. 1–70, volume 2.

[25] Pichler, D., "Extensionality of obligations in Åqvist's system F," Master's thesis, TU Wien (2022).

[26] Prakken, H. and M. Sergot, *Dyadic deontic logic and contrary-to-duty obligations*, in: D. Nute, editor, *Defeasible Deontic Logic*, Kluwer, Dordrecht, 1997 pp. 223–262.

[27] Quine, W. V., *Notes on existence and necessity*, The Journal of Philosophy **40** (1943), pp. 113–127.

[28] Sawasaki, T. and K. Sano, *Term-sequence-dyadic deontic logic*, in: F. Liu, A. Marra, P. Portner and F. V. D. Putte, editors, *Deontic Logic and Normative Systems - 15th International Conference, DEON 2020/21, Munich, Germany [virtual], July 21-24, 2021* (2021), pp. 376–393.

[29] Segerberg, K., *Two-dimensional modal logic*, Journal of Philosophical Logic **2** (1973), pp. 77–96.

[30] Thomason, R. H., *Some completeness results for modal predicate calculi*, in: *Philosophical problems in logic*, Springer, 1970 pp. 56–76.

[31] von Wright, G. H., *Deontic logic*, Mind **60** (1951), pp. 1–15.

Analytic proof theory for Åqvist's system F

Agata Ciabattoni $^{\star\,1}$ Nicola Olivetti $^{\star\star\,2}$ Xavier Parent $^{\star\,1}$
Revantha Rayamanake $^{\star\star\star\,3}$ Dmitry Rozplokhas $^{\star\,1}$

* *Vienna University of Technology, Austria*

** *Aix-Marseille Université, CNRS LIS, France*

*** *University of Groningen, The Netherlands*

Abstract

The key strength of preference-based logics for conditional obligation is their ability to handle contrary-to-duty paradoxes and account for exceptions. Here we investigate Åqvist's system **F**, a well-known logic in this family. **F** has the notable feature that every satisfiable formula has a "best" element. Thus far, the only proof system for **F** was a Hilbert calculus, impeding applications and deeper investigations. We fill this gap, constructing the first analytic calculus for **F**. The calculus possesses good proof-theoretical properties—in particular, cut-elimination, which greatly facilitates proof search. Our calculus is used to provide explanations of logical consequences, as a decision-making tool, and to obtain a preliminary complexity upper bound for **F** (giving a theoretical limit on its automated behavior).

Keywords: Dyadic deontic logic; analytic sequent calculi; hypersequents; system **F**

1 Introduction

This paper deals with so-called preference based dyadic deontic logic, initially put forth by [7,14,28,18]. The syntax contains a conditional obligation operator $\bigcirc(B/A)$, read as"B is obligatory given A". A binary relation ranks the possible worlds in terms of betterness. In that framework, the truth-conditions for $\bigcirc(B/A)$ are phrased in terms of best-antecedents worlds. It has emerged as one of the *de facto* standards for normative reasoning; its key strengths are the ability to handle contrary-to-duty paradoxes [5] and to account for exceptions.

Past research on preference-based dyadic deontic logic has focused on the search for an Hilbert style axiomatization, and on the question of clarifying the correspondence between semantic properties and modal axioms. An overview of the existing findings may be found in, e.g., [11,22]. It is only recently that analytic calculi for these logics have been proposed [24,6]. In an analytic calculus,

[1] {agata, xavier, dmitry}@logic.at

[2] nicola.olivetti@univ-amu.fr

[3] d.r.s.ramanayake@rug.nl

proof search proceeds by step-wise decomposition of the formulas to be proven and this yields practical and theoretical advantages over Hilbert systems. In particular, they can be used to establish important meta-logical properties for the formalized logics (e.g., decidability, complexity and interpolation), and they facilitate the development of automated reasoning methods. The original tool to construct analytic calculi was the sequent calculus; following Gentzen (1933), the key idea was to use the cut rule to establish completeness, and then show elimination (or redundancy) of this rule from derivations to establish analyticity. However, the sequent calculus is not expressive enough to support cut elimination for most logics of interest. Hence various extensions and generalizations have been introduced in the pursuit of analytic proof calculi.

Analytic sequent-style calculi were obtained for two well-known systems proposed by Åqvist [1]: **E** and **G**. For **E** the calculus (called **HE**) was defined in [6], whereas the calculus for **G** appears in [10] (it was in fact introduced for Lewis's *VTA*, to which **G** is equivalent). In this paper we consider **F** which lies between **E** and **G**. It is obtained by supplementing **E** with axiom (D*) that rules out models without a best element (a 'limit'). Obligations in **E** collapse to triviality when there is no best world: if A does not have a best element element then $\bigcirc(B/A)$ holds for any B. One obtains **G**, by extending **F** with the so-called principle of rational monotony [17].

The main contribution of the paper is an analytic calculus **HF** for **F**, leading to a decision procedure and a CoNEXP upper bound, the first complexity bound for this logic. Of Åqvist's three systems, **F** is the most complex in terms of proof theory. **HF** is obtained in a modular way, by adding to (an equivalent version of) **HE** a new rule corresponding to the (D*) axiom. Surprisingly, this rule shares common structural features with the peculiar rule for the calculus for provability logic GL [23]. As in **HE**, the calculus **HF** employs hypersequents to accommodate the extra S5-type modality used to express settledness. The hypersequent framework [2] consists of multiple sequents in parallel, and it can be seen as the minimal extension of Gentzen's sequent framework permitting a cut-free calculus for the logic $S5$ [19,3,16] (itself a sub-logic of **F**). The analyticity of **HF** is established as a consequence of the algorithmic eliminability of the cut rule from derivations (cut-elimination). The proof is intricate and of technical interest. In particular, the presence in the peculiar rule of **HF** of "diagonal formulas" [23] (i.e., formulas that change polarity from conclusion to premises) makes the proof very challenging; even more than in Valentini's cut-elimination proof [25] for GL (see [12,13] for a survey on cut-elimination proofs for GL).

A potential misunderstanding must be cleared up from the start. As in previous work on modal interpretation of conditionals, e.g., [9,20,26,6], we encode maximality by a unary modal operator $\mathcal{B}et$. It is important to realize that by doing so we are not carrying out a reduction of dyadic deontic logic to some bi-modal logic. Indeed the calculus rules for $\mathcal{B}et$ cannot be understood in isolation, and they do not correspond to any known normal or non-normal modality. The $\mathcal{B}et$ operator is not part of the language of **F**, and it is used just in the hypersequent calculus to define suitable rules for the conditional obligation operator.

2 The system F in a nutshell

We present the logic **F**. Its language is defined by the following BNF:

$$A ::= p \in \text{PropVar} \mid \neg A \mid A \to A \mid \Box A \mid \bigcirc(A/A)$$

$\Box A$ is read as "A is settled as true," and $\bigcirc(B/A)$ as "B is obligatory, given A." The Boolean connectives other than $\neg$ and $\to$ are defined as usual. $\Diamond$ is a derived connective, defined as usual (viz. as the dual of $\Box$).

Definition 2.1 **F** consists of any Hilbert system for S5 supplemented with:

$$\bigcirc(B \to C/A) \to (\bigcirc(B/A) \to \bigcirc(C/A)) \qquad \text{(COK)}$$
$$\bigcirc(A/A) \qquad \text{(Id)}$$
$$\bigcirc(C/A \wedge B) \to \bigcirc(B \to C/A) \qquad \text{(Sh)}$$
$$\Box(A \leftrightarrow B) \to (\bigcirc(C/A) \leftrightarrow \bigcirc(C/B)) \qquad \text{(Ext)}$$
$$\bigcirc(B/A) \to \Box\bigcirc(B/A) \qquad \text{(Abs)}$$
$$\Box A \to \bigcirc(A/B) \qquad \text{(O-Nec)}$$
$$\Diamond A \to \neg\bigcirc(\bot/A) \qquad \text{(D}^\star\text{)}$$

F extends **E** with one axiom: (D*). This axiom, which is equivalent to the original axiom, $\Diamond A \to \neg(\bigcirc(B/A) \wedge \bigcirc(\neg B/A))$, rules out the possibility of conflicts between obligations (for consistent, or possible, antecedents).

The notions of derivation and theoremhood are defined in the usual way.

The semantics of **F** can be defined in terms of *preference models*. They are possible-world models equipped with a comparative goodness relation $\succ$ on worlds so that $x \succ y$ can be read as "world x is *better* than world y." Conditional obligation is defined by considering "best" worlds: intuitively, $\bigcirc(B/A)$ holds in a model, if all the best worlds in which A is true also make B true.

Definition 2.2 A preference model is a structure $M = (W, \succ, V)$ $(W \neq \emptyset)$ whose members are called possible worlds, $\succ \subseteq W \times W$, $V : W \to \mathcal{P}(PropVar)$. The evaluation rules for the Boolean connectives are as usual. The evaluation rules for $\Box$ and $\bigcirc$ are defined as follows:

- $M, x \vDash \Box A$ iff $\forall y \in W$ $M, y \vDash A$
- $M, x \vDash \bigcirc(B/A)$ iff $\forall y \in \text{best}(A)$ $M, y \vDash B$

Here $\text{best}(A) = \{y \in W \mid M, y \vDash A \text{ and there is no } z \succ y \text{ such that } M, z \vDash A\}$.

When no confusion arises, we write $x \vDash A$ for $M, x \vDash A$.

The distinctive feature of the semantics for **F** (w.r.t **E**) is that $\succ$ is required to be limited, that is if $\exists x$ s.t. $x \models A$, then $\text{best}(A) \neq \emptyset$. Intuitively, if the set of A-worlds is non-empty, then it has a best element. This assumption validates (D*). Observe that the relation $\succ$ is not assumed to be transitive.

Validity in a model and validity over all models are defined as usual.

For the purpose of the calculi developed subsequently, we introduce the modality $\mathcal{B}et$ that will allow us to represent the "Best" worlds: $M, x \vDash \mathcal{B}et\, A$ iff $\forall y \succ x\ M, y \vDash A$. By this definition, we get $x \in \text{best}(A)$ iff $M, x \vDash A$ and $M, x \vDash \mathcal{B}et\, \neg A$. However, the modality $\mathcal{B}et$ is not part of $\mathcal{L}$.

The following applies:

Theorem 2.3 (Soundness and completeness, [21]) **F** *is sound and complete w.r.t. the class of preference models whose relation* $\succ$ *is limited.*

A few words on the *rationale* behind the limitedness condition. As mentioned, it provides a remedy to the fact that obligations collapse to triviality when there is no best world in a given model. This collapse may arise in two typical situations: when there is an infinite sequence of better and better worlds (see Ex. 2.4 below), and when there is a cycle of betterness (see Ex. 2.5).

Example 2.4 [Starvation, [8]] Let $W = \{x_i : i < \omega\}$. Assume that all the worlds share an infinite number of inhabitants, $\{a_i : i < \omega\}$. In each world x_i, all the individuals whose index is less than or equal to i are relieved and saved from starvation, all the other are left dying. Thus, in x_1, only a_1 is relieved or saved, all the other individuals are starving. In world x_2, only a_1 and a_2 are relieved, all the others are starving, and so on. Suppose the worlds are ranked according to the number of individuals saved from starvation. Then, for all $i < \omega$, $x_{i+1} \succ x_i$. There is no best world. In this model, for all $i < \omega$, (sv_a_i stand for "a_i is saved") $\bigcirc(sv\_a_i/\top)$ and $\bigcirc(\neg sv\_a_i/\top)$, contradicting (D*). Note that $\succ$ has been chosen so as to be transitive. But nothing hinges on it. Indeed, what makes the limitedness condition fail in this model, is that $\succ$ is serial, viz. for all x_i, there is a y such that $y \succ x_i$.

Cycles are usually considered irrational, because they lead to a violation of the principles of transitivity and consistency in decision-making. Nevertheless, empirical studies have revealed that cycles can arise in certain contexts, for instance when the ranking is based on multiple criteria. It is customary to rank the possible worlds based on the number of obligations they violate: the less obligations are violated by a world, the better the world is. This mono-criterion becomes a bi-criterion, if one distinguishes between the obligations issued by an authority P from those issued by an authority Q, and use them separately to rank the possible worlds.

Example 2.5 [Multi-criteria ranking, [27]] Suppose the authorities P and Q issue the commands p_1 and p_2, and q_1 and q_2, respectively. Consider two words x_1 and x_2 such that $x_1 \models p_1 \wedge p_2 \wedge q_1 \wedge \neg q_2$, and $x_2 \models \neg p_1 \wedge p_2 \wedge q_1 \wedge q_2$. We have $x_1 \succ x_2$, since x_1 violates less obligations issued by P than x_2. But $x_2 \succ x_1$, because x_2 violates less obligations issued by Q than x_1. This is a cycle of length 1. In this model, e.g. $\Diamond(p_1 \vee p_2)$, $\bigcirc(p_2/p_1 \vee p_2)$ and $\bigcirc(\neg p_2/p_1 \vee p_2)$, contradicting (D*). As in the previous example, even though $\succ$ has been chosen so as to be transitive, nothing hinges on it. To see why, we use the following variant of $\succ$, putting $x \succ y$ whenever x violates *strictly* less obligations issued by one authority than y does. The outcome is the same. In particular we still have $x_2 \succ x_1$ and $x_1 \succ x_2$. But $\succ$ is no longer transitive (since e.g. $x_2 \not\succ x_2$).

Observe that $\succ$ has been chosen to be total or complete, viz. for all x and y, $x \succ y$ or $y \succ x$. To show that nothing hinges on this property, consider the following variant definition, setting $x \succ y$ whenever the set of obligations issued by one authority that are violated by x is a subset of the set of those violated by y. Suppose the model contains two extra words x_3 and x_4 such that $x_3 \models \neg p_1 \wedge p_2 \wedge \neg q_1 \wedge q_2$, and $x_4 \models p_1 \wedge \neg p_2 \wedge q_1 \wedge \neg q_2$. We have in addition $x_1 \succ x_3$, $x_2 \succ x_4$, $x_1 \succ x_4$, $x_2 \succ x_3$, $x_3 \not\succ x_4$, and $x_4 \not\succ x_3$. Observe that the the outcome is the same even though $\succ$ is not total in this setting.

3 A cut-free hypersequent calculus for F

We introduce the hypersequent calculus **HF** for the logic **F**. **HF** is defined by adding to (a slightly modified[4] version of the) calculus for **E** a new rule ($\mathcal{B}et_F$) corresponding to the (D*) axiom. The resulting calculus extends the hypersequent calculus for $S5$ [3,16] with left and right rules for the dyadic obligation, and two rules for $\mathcal{B}et$ (the **HE** calculus for **E** had only one).

Introduced in [19] to define a cut-free calculus for $S5$, hypersequents consist of sequents working in parallel.

Definition 3.1 [2] A *hypersequent* is a multiset $\Gamma_1 \Rightarrow \Pi_1 \mid \ldots \mid \Gamma_n \Rightarrow \Pi_n$ where, for all $i = 1, \ldots, n$, $\Gamma_i \Rightarrow \Pi_i$ is a multisets-based sequent, called a *component* of the hypersequent.

The hypersequent calculus **HF** is presented in Definition 3.2. **HF** consists of initial hypersequents (i.e., axioms), logical/modal/deontic and structural rules. The latter are divided into *internal* and *external rules*. **HF** incorporates the sequent calculus for the modal logic S4 as a sub-calculus and adds an additional layer of information by considering a single sequent to live in the context of hypersequents. Hence all the axioms and rules of **HF** (but the external structural rules) are obtained by adding to each sequent a context G or H, representing a (possibly empty) hypersequent. For instance, the (hypersequent version of the) axioms are $\Gamma, p \Rightarrow \Delta, p \mid G$. The external structural rules include ext. weakening (ew) and ext. contraction (ec) (see Fig. 1). These behave like weakening and contraction over whole hypersequent components. The hypersequent structure opens the possibility to define new such rules that allow the "exchange of information" between different sequents. These type of rules increases the expressive power of hypersequent calculi compared to sequent calculi, enabling the definition of cut-free calculi for logics that seem to escape a cut-free sequent formulation (e.g., $S5$). An example of external structural rule is the ($s5$) rule in [16] (reformulated as ($s5'$) in Fig. 1 to account for the presence of $\bigcirc$), that allows the peculiar axiom of $S5$ to be derived.

The rules in Fig. 1 and 2 make use of the following notation:

$$\Sigma^{\Box} = \{\Box B : \ \Box B \in \Sigma\} \quad \Sigma^{O} = \{\bigcirc(C/D) : \bigcirc(C/D) \in \Sigma\} \quad \Sigma^{\Box,O} = \Sigma^{\Box}, \Sigma^{O}$$

[4] We employ a version of the rule ($\mathcal{B}et$) that contains exactly one formula on its LHS, see Remark 3.10.

$$\frac{G}{G \mid \Gamma \Rightarrow \Pi}\,(ew) \qquad \frac{G \mid \Gamma \Rightarrow \Pi \mid \Gamma \Rightarrow \Pi}{G \mid \Gamma \Rightarrow \Pi}\,(ec) \qquad \frac{G \mid \Gamma^{\Box}, \Gamma^{O}, \Gamma' \Rightarrow \Pi'}{G \mid \Gamma \Rightarrow\ \mid \Gamma' \Rightarrow \Pi'}\,(s5')$$

Fig. 1. External structural rules

Also, for any set $\mathbb{D}$ of formulae, define $\mathcal{B}et\,\mathbb{D}$ as the set $\{\mathcal{B}et\,D \mid D \in \mathbb{D}\}$.

Definition 3.2 The hypersequent calculus **HF** consists of the hypersequent version of Gentzen LK sequent calculus for propositional classical logic, the external structural rules in Fig. 1, and the modal and deontic rules in Fig. 2.

Lemma 3.3 *The rules* $(\bigcirc R), (\mathcal{B}et), (\Box R)$ *and* (Bet_F) *are equivalent in* **HF** *to their version* $(\bigcirc R)^*, (\mathcal{B}et)^*, (\Box R)^*$, *and* $(Bet_F)^*$ *without the internal contexts* Γ *and* $\Gamma^{\Box,O}$.

Proof. One direction is trivial. For the other direction, consider the case of $(\bigcirc R)$ (the other cases are similar), and the following proof

$$\dfrac{\dfrac{\dfrac{\dfrac{\Gamma^{\Box,O}, A, \mathcal{B}et\,\neg A \Rightarrow B \mid G}{A, \mathcal{B}et\,\neg A \Rightarrow B \mid \Gamma^{\Box,O} \Rightarrow\ \mid G}\,(s5')}{\Rightarrow \bigcirc(B/A) \mid \Gamma^{\Box,O} \Rightarrow\ \mid G}\,(\bigcirc \mathrm{R})^*}{\Gamma \Rightarrow \bigcirc(B/A) \mid \Gamma \Rightarrow \bigcirc(B/A) \mid G}\,(\mathrm{w})}{\Gamma \Rightarrow \bigcirc(B/A) \mid G}\,(\mathrm{ec})$$

□

Remark 3.4 The $(\mathcal{B}et_F)$ rule corresponds to the condition of limitedness of the betterness relation. A natural way to express this condition as a hypersequent rule is

$$\frac{G \mid \Gamma^{\Box,O}, \mathcal{B}et\,A \Rightarrow A}{G \mid \Gamma \Rightarrow A}$$

The upper sequent encodes the fact that in an arbitrary model $best(\neg A) = \emptyset$ (i.e. for any world x, if $y \vDash A$ for all $y \succ x$, then $x \vDash A$ also). The limitedness condition states that this can only happen if there is no world where $\neg A$ holds. The lower sequent encodes this fact. However, the addition of this rule to the calculus **HE** is not enough to obtain a complete *cut-free* calculus. The same holds for the one premise version of $(\mathcal{B}et_F)$, viz

$$\frac{G \mid \Gamma^{\Box,O}, \mathcal{B}et\,A \Rightarrow A}{G \mid \Gamma \Rightarrow \mathcal{B}et\,A}$$

In a calculus with this sole rule, the following formula (where a, b, and c are propositional variables) cannot be derived without using the cut rule:

$$\bigcirc(b \wedge \neg c/a) \wedge \bigcirc(a \wedge c/b) \rightarrow \bigcirc(\bot/a)$$

Ex. 3.9 shows how $(\mathcal{B}et_F)$ enables to get a cut free derivation of this formula.

$$\frac{\Gamma^{\Box,O}, A, \mathcal{B}et\,\neg A \Rightarrow B \,|\, G}{\Gamma \Rightarrow \Delta, \bigcirc(B/A) \,|\, G}\ (\bigcirc R) \qquad \frac{\Gamma^{\Box,O}, B \Rightarrow A \,|\, G}{\Gamma, \mathcal{B}et\, B \Rightarrow \Delta, \mathcal{B}et\, A \,|\, G}\ (\mathcal{B}et)$$

$$\frac{\Gamma^{\Box,O} \Rightarrow A \,|\, G}{\Gamma \Rightarrow \Delta, \Box A \,|\, G}\ (\Box R) \qquad \frac{\Gamma, A \Rightarrow \Delta \,|\, G}{\Gamma, \Box A \Rightarrow \Delta \,|\, G}\ (\Box L)$$

$$\frac{\{\Gamma^{\Box,O}, \mathcal{B}et\,\mathbb{D}, \mathcal{B}et\, B \Rightarrow D_i \,|\, G\}_{D_i \in \mathbb{D}} \qquad \Gamma^{\Box,O}, \mathcal{B}et\,\mathbb{D}, \mathcal{B}et\, B \Rightarrow B \,|\, G}{\Gamma, \mathcal{B}et\,\mathbb{D} \Rightarrow \Delta, \mathcal{B}et\, B \,|\, G}\ (\mathcal{B}et_F)$$

$$\frac{\Gamma, \bigcirc(B/A) \Rightarrow \Delta, A \,|\, G \quad \Gamma, \bigcirc(B/A) \Rightarrow \Delta, \mathcal{B}et\,\neg A \,|\, G \quad \Gamma, \bigcirc(B/A), B \Rightarrow \Delta \,|\, G}{\Gamma, \bigcirc(B/A) \Rightarrow \Delta \,|\, G}\ (\bigcirc L)$$

Fig. 2. Deontic and modal rules

A *derivation* in **HF** is a (possibly infinite) tree obtained by applying the rules bottom up. A *proof* $\mathcal{D}$ is a finite derivation whose leaves are axioms.

The soundness of **HE** is proved with respect to preference models. Although we can interpret a hypersequent H directly into the semantics, it is easier (and more readable) to interpret it as a formula $I(H)$ of the extended language $\mathcal{L} + \mathcal{B}et$. Then validity of $I(H)$ is defined as usual. We now show the validity of this formula whenever H is provable.

Theorem 3.5 *If there is a proof in* **HF** *of* $H := \Gamma_1 \Rightarrow \Pi_1 \,|\, \ldots \,|\, \Gamma_n \Rightarrow \Pi_n$*, then* $I(H) := \Box(\bigwedge \Gamma_1 \rightarrow \bigvee \Pi_1) \vee \ldots \vee \Box(\bigwedge \Gamma_n \rightarrow \bigvee \Pi_n)$ *is valid.*

Proof. We only need to show the soundness of the new rule $(\mathcal{B}et_F)$. Soundness of the other rules w.r.t. **F** follows from their soundness w.r.t. the weaker logic **E** proved in [6]. This includes the $(\mathcal{B}et)$ rule of **HF**, which is a weakened version of the homonymous rule of **HE**. It is enough to establish soundness for the simplified version of $(\mathcal{B}et_F)$ without internal contexts since the original version can be obtained via its combination with sound structural rules (Lemma 3.3). Suppose all the premises of $(\mathcal{B}et_F)$ are valid but not the conclusion. Thus, for some model M whose relation $\succ$ is limited and some world w in it, $w \nvDash \Box(\bigwedge \mathcal{B}et\,\mathbb{D} \rightarrow \mathcal{B}et\, B) \vee I(G)$. Thus $w \nvDash I(G)$ and therefore (1) $I(G)$ does not hold in any world–$I(G)$ is a disjunction of formulas prefixed with $\Box$, and gets the same truth-value in all worlds. Also, for some world x, $x \nvDash \mathcal{B}et\, B$. Therefore, there exists a world $y \succ x$ such that $y \nvDash B$, viz. $y \vDash \neg B$ and so $y \vDash \neg(B \wedge \bigwedge_{D_i \in \mathbb{D}} D_i)$. By the limitedness condition, there exists a world z in M that belongs to best$(\neg(B \wedge \bigwedge_{D_i \in \mathbb{D}} D_i))$, i.e. (2) $z \nvDash B \wedge \bigwedge_{D_i \in \mathbb{D}} D_i$ and (3) for all $u \succ z$, $u \vDash B \wedge \bigwedge_{D_i \in \mathbb{D}} D_i$. By (2), (4) either $z \nvDash B$ or $z \nvDash D_j$ for some $D_j \in \mathbb{D}$. Consider the second case. From the opening assumption (using the left-most premise, with D_j on the right) and (1), one gets $z \models \bigwedge \mathcal{B}et\,\mathbb{D} \wedge \mathcal{B}et\, B \rightarrow D_j$. By contraposition, one gets that either $z \not\models \mathcal{B}et\, B$ or $z \not\models \mathcal{B}et\, D_k$ for some $D_k \in \mathbb{D}$. This contradicts (3). The case when $z \nvDash B$ in (4) is handled analogously (now using the right-most premise, with B on the right hand side of $\Rightarrow$). $\Box$

Lemma 3.6 *For every formula A: $A \Rightarrow A$ is derivable in* **HF**.

Proof. Standard induction on the complexity of A. □

Theorem 3.7 (Completeness with cut) *Each theorem of* **F** *has a proof in* **HF** *with the addition of the cut rule:*

$$\dfrac{G \,|\, \Gamma, A \Rightarrow \Delta \qquad H \,|\, \Sigma \Rightarrow \Pi, A}{G \,|\, H \,|\, \Gamma, \Sigma \Rightarrow \Delta, \Pi}\ (cut)$$

Proof. (D*) axiom (with $\Diamond$ rewritten as $\neg\Box\neg$) can be derived as follows:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{A \Rightarrow A}{\Rightarrow \neg A, A}{\scriptstyle(\neg\mathrm{R})} \qquad \dfrac{\dfrac{\dfrac{\dfrac{A \Rightarrow A}{\Rightarrow \neg A, A}{\scriptstyle(\neg\mathrm{R})} \qquad \mathcal{B}et\,\neg A \Rightarrow \mathcal{B}et\,\neg A \qquad \dfrac{}{\bot \Rightarrow}{\scriptstyle(\bot\mathrm{L})}}{\bigcirc(\bot/A), \mathcal{B}et\,\neg A \Rightarrow \neg A}{\scriptstyle(\bigcirc\mathrm{L})+(\mathrm{w})}}{\bigcirc(\bot/A) \Rightarrow \mathcal{B}et\,\neg A}{\scriptstyle(\mathcal{B}et_{\mathrm{F}})}}{} \qquad \dfrac{}{\bot \Rightarrow}{\scriptstyle(\bot\mathrm{L})}}{\bigcirc(\bot/A) \Rightarrow \neg A}{\scriptstyle(\bigcirc\mathrm{L})+(\mathrm{w})}}{\bigcirc(\bot/A) \Rightarrow \Box\neg A}{\scriptstyle(\Box\mathrm{R})}}{\neg\Box\neg A \Rightarrow \neg \bigcirc (\bot/A)}{\scriptstyle(\neg\mathrm{L})+(\neg\mathrm{R})}}{\Rightarrow \neg\Box\neg A \ \rightarrow\ \neg \bigcirc (\bot/A)}{\scriptstyle(\rightarrow\mathrm{R})}$$

(*Nec*) and all the remaining axioms are provable in **HF** (without using ($\mathcal{B}et_F$) or (*cut*)), while modus ponens requires (*cut*). □

Example 3.8 The Kantian principle "ought implies can" $\bigcirc(B/A) \rightarrow (\Diamond A \rightarrow \Diamond(A \wedge B))$ holds in **F**, as shown by the following **HF** proof (we omit straightforward subderivations of propositional tautologies in the leaves)

$$\dfrac{\dfrac{\dfrac{\Rightarrow \neg A, A \qquad \dfrac{\dfrac{\Rightarrow \neg A, A \qquad \mathcal{B}et\,\neg A \Rightarrow \mathcal{B}et\,\neg A \qquad \dfrac{\neg(A \wedge B), B \Rightarrow \neg A}{\Box\neg(A \wedge B), B \Rightarrow \neg A}{\scriptstyle(\Box\mathrm{L})}}{\bigcirc(B/A), \Box\neg(A \wedge B), \mathcal{B}et\,\neg A \Rightarrow \neg A}{\scriptstyle(\bigcirc\mathrm{L})+(\mathrm{w})}}{\bigcirc(B/A), \Box\neg(A \wedge B) \Rightarrow \mathcal{B}et\,\neg A}{\scriptstyle(\mathcal{B}et_{\mathrm{F}})} \qquad \dfrac{\neg(A \wedge B), B \Rightarrow \neg A}{\Box\neg(A \wedge B), B \Rightarrow \neg A}{\scriptstyle(\Box\mathrm{L})}}{\bigcirc(B/A), \Box\neg(A \wedge B) \Rightarrow \neg A}{\scriptstyle(\bigcirc\mathrm{L})+(\mathrm{w})}}{\bigcirc(B/A), \Box\neg(A \wedge B) \Rightarrow \Box\neg A}{\scriptstyle(\Box\mathrm{R})}}{\Rightarrow \bigcirc(B/A) \ \rightarrow\ (\neg\Box\neg A \ \rightarrow\ \neg\Box\neg(A \wedge B))}{\scriptstyle(\rightarrow\mathrm{R})\mathrm{x2}+(\neg\mathrm{L})+(\neg\mathrm{R})}$$

Example 3.9 A derivation in **HF** of the formula in Remark 3.4 is as follows. First, we eliminate connectives and modalities in a natural fashion (we omit the premises of the ($\bigcirc L$) rule applications that are propositional tautologies):

$$
\dfrac{\dots \qquad \dfrac{\dots \qquad \dfrac{(1) \qquad (2)}{\bigcirc(b \wedge \neg c/a), \bigcirc(a \wedge c/b), \mathcal{B}et\,\neg a \Rightarrow \mathcal{B}et\,\neg b}\,(\mathcal{B}et_{\mathrm{F}}) \qquad \dots}{\bigcirc(b \wedge \neg c/a), \bigcirc(a \wedge c/b), a, \mathcal{B}et\,\neg a, b \wedge \neg c \Rightarrow \bot}\,(\bigcirc\mathrm{L})+(\mathrm{w})}{\dfrac{\dfrac{\bigcirc(b \wedge \neg c/a), \bigcirc(a \wedge c/b), a, \mathcal{B}et\,\neg a \Rightarrow \bot}{\bigcirc(b \wedge \neg c/a), \bigcirc(a \wedge c/b) \Rightarrow \bigcirc(\bot/a)}\,(\bigcirc\mathrm{R})}{\Rightarrow \bigcirc(b \wedge \neg c/a) \wedge \bigcirc(a \wedge c/b) \;\to\; \bigcirc(\bot/a)}\,(\to\mathrm{R})+(\wedge\mathrm{L})}\,(\bigcirc\mathrm{L})+(\mathrm{w})
$$

The sequent $\bigcirc(b \wedge \neg c/a), \bigcirc(a \wedge c/b), \mathcal{B}et\,\neg a \Rightarrow \mathcal{B}et\,\neg b$ can be derived by applying the $(\mathcal{B}et_F)$ rule on both Bet-formulas leading to the two premises (1) $O(b \wedge \neg c/a), O(a \wedge c/b), \mathcal{B}et\,\neg a, \mathcal{B}et\,\neg b \Rightarrow \neg a$ and (2) $O(b \wedge \neg c/a), O(a \wedge c/b), \mathcal{B}et\,\neg a, \mathcal{B}et\,\neg b \Rightarrow \neg b$, both of which can be proved by applying $(\bigcirc L)$ once again.

Remark 3.10 The calculus **HF** is obtained by extending the hypersequent calculus for S5 with suitable rules for $\bigcirc$, making use of the auxiliary modality $\mathcal{B}et$. It is easy to see that $\bigcirc$ can be actually defined in the language with $\Box$ and $\mathcal{B}et$ in the sense that we have the following semantic equivalence

$$\bigcirc(B/A) \equiv \Box((A \wedge \mathcal{B}et\,\neg A) \to B)$$

It is possible to obtain a proof system for F by treating $\bigcirc(B/A)$ as an abbreviation. However, we have not done so for two reasons. First of all, this makes it possible to study $\bigcirc$ even in the $\Box$-free fragment of **F**. Moreover, we have a complete calculus for the $\Box$-free fragment of **F** (and this entails that the addition of $\Box$ is conservative). The second reason is that the way the $\mathcal{B}et$-modality is treated in the calculus **HF** does not correspond to any known normal or non-normal modal logic. For instance, it is easily seen that the $(\mathcal{B}et)$-rule does not allow us to derive standard axioms of the modal logic K, like $(\mathcal{B}et\,A \wedge \mathcal{B}et\,B) \to \mathcal{B}et(A \wedge B)$. Therefore the rules for $\mathcal{B}et$ *are not complete* with respect to its semantics for proving arbitrary sequents in the language with $\mathcal{B}et$ (as opposed to sequents containing formulas in the language of **F**). Also observe that the following rule (cf. Remark 3.4) is valid:

$$\frac{\mathcal{B}et\,A \to A}{\mathcal{B}et\,A}$$

This rule is not valid in K, but it is in GL (see, e.g. [25]). In conclusion, **HF** is not the combination of two existing calculi, one for $S5$ and one for $\mathcal{B}et$.

3.1 Cut-elimination

The completeness proof of **HF** makes use of the cut rule. Here we give a constructive proof that cut can be *eliminated* from **HF**+*cut* proofs. This result (cut-elimination) is typically proved by stepwise applications of permutation and principal reductions. The former shifts a cut one step upwards in either the left premise or the right premise. Following repeated applications, the situation is reached of a cut in which the cut-formula is principal (i.e. created by the rule immediately above it) in both premises. The principal reduction is now used to replace that cut with cuts on proper subformulas. An appeal to (transfinite) induction ultimately yields a cut-free proof.

The cut-elimination proof for **HF** + *cut* is not an easy adaptation of the corresponding result for **HE**. Indeed, the presence in (Bet_F) of a formula $\mathcal{B}et\, B$ that changes polarity from conclusion to premises, makes the principal reduction step even more involved than in the modal logic of provability GL.

The immediate corollary of cut-elimination is (a relaxed form of) the *subformula property*: every formula in a **HF** proof is a subformula (possibly negated and under the scope of $\mathcal{B}et$) of the end-formula.

Roadmap of the proof: To reduce the complexity of the cut on a → or ¬-formula we exploit the invertibility of its introduction rules (Lemma 3.11) and the usual principal reduction steps. Invertibility does not work for formulas of the form $\Box A$, $\mathcal{B}et\, A$, and $\bigcirc(B/A)$ so cuts have to be shifted upward till the cut-formula is introduced. The first challenge, already witnessed in **HE** is that the $(\Box R)$, $(\bigcirc R)$ and $(\mathcal{B}et)$ (as well as $(\mathcal{B}et_F)$) rules cannot be shifted below *every* cut: only those involving hypersequents of a certain "good" shape. Therefore a specific reduction strategy for lifting uppermost cuts is required: first over the premise in which the cut formula appears on the right (Lemma 3.15) and then, when a rule introducing the cut formula is reached (and in this case the sequent has a "good" shape), shifting the cut upwards over the other premise (Lemma 3.14) and then applying the principal reduction. This last reduction step is "standard" for $\Box$, $\bigcirc$-formulas and $\mathcal{B}et$ formulas introduced on both sides by $(\mathcal{B}et)$ (Lemma 3.12), while when $\mathcal{B}et$ formulas are introduced by $(\mathcal{B}et_F)$ a sophisticated argument inspired by the cut-elimination proof for the logic GL [25] is used. This is the second, and main, challenge in proving cut-elimination for **HF**. Note that the hypersequent structure itself does not necessitate major changes: the $(s5')$ rule permits permutation with cuts of a "good" shape, and to handle (ec) we consider the hypersequent version of the multicut: cut one component against (possibly) many components.

Notation and Terminology. The *length* $|\mathcal{D}|$ of an **HF** proof $\mathcal{D}$ is (the maximal number of applications of inference rules) +1 occurring on any branch of d. The *complexity* $\lceil A \rceil$ of a formula A is defined as: $\lceil A \rceil = 1$ if A is atomic, $\lceil \neg A \rceil = \lceil A \rceil + 1$, $\lceil A \to B \rceil = \lceil A \rceil + \lceil B \rceil + 1$, $\lceil \mathcal{B}et\, A \rceil = \lceil A \rceil + 1$, $\lceil \Box A \rceil = \lceil A \rceil + 1$, and $\lceil \bigcirc(A/B) \rceil = \lceil A \rceil + \lceil B \rceil + 3$. The *cut rank* $\rho(\mathcal{D})$ of $\mathcal{D}$ is the maximal complexity of cut formulas in $\mathcal{D}$, so $\rho(\mathcal{D}) = 0$ if $\mathcal{D}$ is cut-free. We use A^n (resp. Γ^n) to indicate n occurrences of A (resp. of Γ).

The rules of the classical propositional connectives remain invertible.

Lemma 3.11 (invertible connectives) *Every* **HF** *proof $\mathcal{D}$ of a hypersequent containing a formula $\neg A$ (resp. $A \to B$), can be transformed into a proof $\mathcal{D}'$ of the same hypersequent ending in an introduction rule for $\neg A$ (resp. $A \to B$) such that $\rho(\mathcal{D}') \leq \rho(\mathcal{D})$.*

As shown below, any cut whose cut formula is immediately introduced in left and right premise can be replaced by smaller cuts. While for compound formulas not introduced by the rule $(\mathcal{B}et_F)$ the transformation is easy, this last case requires Lemma 3.13.

Lemma 3.12 (reduce principal cuts) *Let A be a compound formula and $\mathcal{D}_l$ and $\mathcal{D}_r$ be* **HF** *proofs such that $\rho(\mathcal{D}_l) < \lceil A \rceil$ and $\rho(\mathcal{D}_r) < \lceil A \rceil$, and*

(i) *$\mathcal{D}_l$ is a proof of $G \mid \Gamma, A \Rightarrow \Delta$ ending in a rule introducing A*

(ii) *$\mathcal{D}_r$ is a proof of $H \mid \Sigma \Rightarrow A, \Pi$ ending in a rule introducing A*

There is a transformation of these proofs into a **HF** *proof of $G \mid H \mid \Gamma, \Sigma \Rightarrow \Delta, \Pi$ with $\rho(\mathcal{D}) < \lceil A \rceil$.*

Proof. We discuss the only non-standard case: $A = \mathcal{B}et\, B$, and use a simplified version of the rules without internal contexts (cf. Lemma 3.3).

Assume that $\mathcal{B}et\, B$ is introduced by two $(\mathcal{B}et)$ rules as in

$$\dfrac{\dfrac{G \mid B \Rightarrow C}{G \mid \Sigma, \mathcal{B}et\, B \Rightarrow \mathcal{B}et\, C, \Pi}\,{\scriptstyle(\mathcal{B}et)} \qquad \dfrac{H \mid D \Rightarrow B}{H \mid \Gamma, \mathcal{B}et\, D \Rightarrow \mathcal{B}et\, B, \Delta}\,{\scriptstyle(\mathcal{B}et)}}{G \mid H \mid \Gamma, \Sigma, \mathcal{B}et\, D \Rightarrow \mathcal{B}et\, C, \Delta, \Pi}\,{\scriptstyle(\mathrm{cut})}$$

the above cut is replaced by

$$\dfrac{\dfrac{G \mid B \Rightarrow C \qquad H \mid D \Rightarrow B}{G \mid H \mid D \Rightarrow C}\,{\scriptstyle(\mathrm{cut})}}{G \mid H \mid \Gamma, \Sigma, \mathcal{B}et\, D \Rightarrow \mathcal{B}et\, C, \Delta, \Pi}\,{\scriptstyle(\mathcal{B}et)}$$

Assume that $\mathcal{B}et\, B$ is introduced on the right hand side by $(\mathcal{B}et_F)$ as in

$$\dfrac{\dfrac{G \mid B \Rightarrow C}{G \mid \Sigma, \mathcal{B}et\, B \Rightarrow \mathcal{B}et\, C, \Pi}\,{\scriptstyle(\mathcal{B}et)} \qquad \dfrac{\{H \mid \mathcal{B}et\, \mathbb{D}, \mathcal{B}et\, B \Rightarrow D_i\}_{1 \leq i \leq n} \qquad H \mid \mathcal{B}et\, \mathbb{D}, \mathcal{B}et\, B \Rightarrow B}{H \mid \Gamma, \mathcal{B}et\, \mathbb{D} \Rightarrow \mathcal{B}et\, B, \Delta}\,{\scriptstyle(\mathcal{B}et_{\mathrm{F}})}}{G \mid H \mid \Gamma, \Sigma, \mathcal{B}et\, \mathbb{D} \Rightarrow \mathcal{B}et\, C, \Delta, \Pi}\,{\scriptstyle(\mathrm{cut})}$$

This case cannot be simply handled by cutting the premises of $(\mathcal{B}et)$ and $(\mathcal{B}et_F)$, because of the additional formulas $\mathcal{B}et\, B$ *on the left* appearing in the premises of $(\mathcal{B}et_F)$. The strategy is to apply Lemma 3.13 to all premises of (Bet_F) to get proofs, with cut-rank $< \lceil \mathcal{B}et\, B_N \rceil$, of the same hypersequents but with $\mathcal{B}et\, B$ on the left removed. Hence we get

$$\dfrac{\dfrac{\{H \mid \mathcal{B}et\, \mathbb{D} \Rightarrow D_i\}_{1 \leq i \leq n}}{\{G \mid H \mid \mathcal{B}et\, \mathbb{D}, \mathcal{B}et\, C \Rightarrow D_i\}_{1 \leq i \leq n}}\,{\scriptstyle(\mathrm{ew})+(\mathrm{w})} \qquad \dfrac{\dfrac{G \mid B \Rightarrow C \qquad H \mid \mathcal{B}et\, \mathbb{D} \Rightarrow B}{G \mid H \mid \mathcal{B}et\, \mathbb{D} \Rightarrow C}\,{\scriptstyle(\mathrm{cut})}}{G \mid H \mid \mathcal{B}et\, \mathbb{D}, \mathcal{B}et\, C \Rightarrow C}\,{\scriptstyle(\mathrm{w})}}{G \mid H \mid \Gamma, \Sigma, \mathcal{B}et\, \mathbb{D} \Rightarrow \mathcal{B}et\, C, \Delta, \Pi}\,{\scriptstyle(\mathcal{B}et_{\mathrm{F}})}$$

□

The following lemma allows us to remove any application of $\mathcal{B}et\, B$ formulas that appear on the left hand side of the $(\mathcal{B}et_F)$ rule, via suitable cuts on B. Its proof is inspired by Valentini's cut-elimination argument for provability logic GL [25] where the corresponding lemma provides a constructive proof of Löb's theorem in GL. It requires indeed to perform global transformations: tracing bottom up from all the premises of (Bet_F) all the *occurrences* (ancestors) of the $\mathcal{B}et\,B$ formulas and substituting them with suitable formulas, taking care that the resulting proof is still correct. The tracing works as follows: we denote by $\mathcal{B}et\,B^*$ a *decorated occurrence* of $\mathcal{B}et\,B$. Starting with a hypersequent with one decorated occurrence of $\mathcal{B}et\,B$, we propagate the decoration through the proof to all formulas $\mathcal{B}et\,B$ which are in a predecessor relation [5] with $\mathcal{B}et\,B^*$. The tracing terminates at an *upper sequent* that is either (a) an axiom $\Gamma, \mathcal{B}et\,B^*, p \Rightarrow p, \Delta$, or the conclusion (b) of an internal/external weakening or of a rule with weakening built in (i.e., $(\Box R)$, and $(\bigcirc R)$), or (c) of $(\mathcal{B}et)$. In the following, for $\mathbb{B} = B_1, \ldots, B_N$, we write $\mathbb{B}_j$ to denote $\mathbb{B} \setminus B_j$.

Lemma 3.13 *Let $\mathcal{D}_1, \ldots, \mathcal{D}_n$ be the following* $\mathbf{HF} + cut$ *proofs of the premises of a $(\mathcal{B}et_F)$ rule instance.*

$$\begin{array}{ccc} \mathcal{D}_1 & & \mathcal{D}_n \\ G \mid \mathcal{B}et\,\mathbb{B} \Rightarrow B_1 & \ldots & G \mid \mathcal{B}et\,\mathbb{B} \Rightarrow B_n \end{array}$$

Suppose that N satisfies $1 \leq N \leq n$, and $\rho(\mathcal{D}_i) < \lceil \mathcal{B}et\,B_N \rceil$ for each i $(1 \leq i \leq n)$. There is a transformation of these proofs into a $\mathbf{HF} + cut$ *proof with cut-rank $< \lceil \mathcal{B}et\,B_N \rceil$ of $G \mid \mathcal{B}et\,\mathbb{B}_N \Rightarrow B_i$ for each i.*

Proof. Observe that the lemma is easy to prove using cut if we remove the requirement that the resulting proof has cut-rank $< \lceil \mathcal{B}et\,B_N \rceil$ (apply $(\mathcal{B}et_F)$ to $\mathcal{D}_1, \ldots, \mathcal{D}_n$ to get $\mathcal{B}et\,\mathbb{B}_N \Rightarrow \mathcal{B}et\,B_N$, then apply cut with the latter to each $\mathcal{D}_i$). To reduce clutter we omit the external contexts and the modal internal contexts as they do not play a role in the argument (cf. Lem. 3.3 for the latter).

Trace $\mathcal{B}et\,B_N$ upwards in each $\mathcal{D}_i$ (we indicate with $\mathcal{B}et\,B_N^*$ its decorated version) until the upper sequents ((a)-(c) above) introducing $\mathcal{B}et\,B_N^*$ are encountered. Define the *depth of* $\mathcal{B}et\,B_N^*$ for a proof ending in a $(\mathcal{B}et_F)$ rule as the total number (over all of its premises) of $(\mathcal{B}et_F)$ rules that contain the decorated formula $\mathcal{B}et\,B_N^*$. Note that $\mathcal{B}et\,B_N^*$ can only appear on the LHS of sequents. We prove the claim by induction on the depth K of $\mathcal{B}et\,B_N^*$ in the premises $\mathcal{D}_1, \ldots, \mathcal{D}_n$.

Inductive case. Suppose that the depth $K > 0$. In that case there must be a nearest $(\mathcal{B}et_F)$ rule above the root of some $\mathcal{D}_i$ of the form

$$\frac{\{\mathcal{B}et\,B_N^*, \mathcal{B}et\,\mathbb{D} \Rightarrow D_i\}_{1 \leq i \leq I} \qquad \mathcal{B}et\,B_N^*, \mathcal{B}et\,\mathbb{D} \Rightarrow B_N}{\mathcal{B}et\,B_N^*, \mathcal{B}et\,\mathbb{D}_I \Rightarrow \mathcal{B}et\,D_I}\ (\mathcal{B}et_F) \qquad (1)$$

Each premise of the above is one $(\mathcal{B}et_F)$ rule away from the root of $\mathcal{D}_i$ and so the depth of $\mathcal{B}et\,B_N^*$ in (1) must be $< K$. Hence we can apply IH to obtain proofs with cut-rank $< \lceil \mathcal{B}et\,B_N \rceil$ of $\mathcal{B}et\,\mathbb{D} \Rightarrow D_i$ for every i $(1 \leq i \leq I)$.

Let $\mathcal{D}'_i$ be obtained from $\mathcal{D}_i$ by replacing the subproof concluding (1) with

[5] This is the familiar parametric ancestor relation of [4] in the setting of hypersequents.

$$\dfrac{\dfrac{\{\mathcal{B}et\,\mathbb{D} \Rightarrow D_i\}_{1\leq i\leq I}}{\mathcal{B}et\,\mathbb{D}_I \Rightarrow \mathcal{B}et\,D_I}\,(\mathcal{B}et_F)}{\mathcal{B}et\,B^*_N, \mathcal{B}et\,\mathbb{D}_I \Rightarrow \mathcal{B}et\,D_I}\,(\mathrm{w})$$

Since we replaced a $(\mathcal{B}et_F)$ rule between the root and the upper sequent with a weakening on $\mathcal{B}et\,B^*_N$, it follows that the depth of $\mathcal{B}et\,B^*_N$ in $\mathcal{D}_1, \ldots, \mathcal{D}'_i, \ldots, \mathcal{D}_n$ (n elements) is $< K$. From the IH we obtain proofs of $\mathcal{B}et\,\mathbb{B}_N \Rightarrow B_i$ with cut-rank $< \lceil \mathcal{B}et\,B_N \rceil$ for every i so the claim is proved.

Base case $K = 0$: there are no $(\mathcal{B}et_F)$ rule instances involving $\mathcal{B}et\,B^*_N$. In this case, when replacing the decorated formula $\mathcal{B}et\,B^*_N$ with suitable formulas, only the upper sequents arising from applications of $(\mathcal{B}et)$ (i.e. case (c)) need some care. We illustrate the proof strategy with a concrete example. See the Appendix for full details.

Suppose that the following upper sequents occur in $\mathcal{D}_1, , \ldots \mathcal{D}_n$.

$$\dfrac{B_N \Rightarrow C}{\mathcal{B}et\,B^*_N \Rightarrow \mathcal{B}et\,C} \qquad \dfrac{B_N \Rightarrow D}{\mathcal{B}et\,B^*_N \Rightarrow \mathcal{B}et\,D} \qquad \dfrac{B_N \Rightarrow B_N}{\mathcal{B}et\,B^*_N \Rightarrow \mathcal{B}et\,B_N}$$

Replace $\mathcal{B}et\,B^*_N$ with $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,C, \mathcal{B}et\,D$ throughout $\mathcal{D}_1, , \ldots \mathcal{D}_n$. The first two upper sequents above become quasi-axioms (cf. Lem. 3.6) $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,C, \mathcal{B}et\,D \Rightarrow \mathcal{B}et\,C$ and $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,C$, $\mathcal{B}et\,D \Rightarrow \mathcal{B}et\,D$, respectively. The third upper sequent now looks like $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,C, \mathcal{B}et\,D \Rightarrow \mathcal{B}et\,B_N$; the latter sequent is provable by applying $(\mathcal{B}et_F)$ to the conclusions of $\mathcal{D}_1, \ldots, \mathcal{D}_n$ (followed by some weakening). In this way we obtain proofs of $(*)$ $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,C, \mathcal{B}et\,D \Rightarrow B_i$ for each i. Now, by two applications of cut on B_N (with the premises $B_N \Rightarrow C$ and $B_N \Rightarrow D$ that appeared in the upper sequents indicated above), we also get $(**)$ $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,C, \mathcal{B}et\,D \Rightarrow C$ and $(***)$ $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,C, \mathcal{B}et\,D \Rightarrow D$. An application of $(\mathcal{B}et_F)$ with premises $(*) - (***)$ leads to a proof of $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,D \Rightarrow \mathcal{B}et\,C$.

Next, replace $\mathcal{B}et\,B^*_N$ with $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,D$ throughout the original $\mathcal{D}_1, \ldots \mathcal{D}_n$ (once again, as in the previous paragraph, the replacements are made in the *original proofs*; this is a feature of the transformation that is seen also in the next paragraph). Then $\mathcal{B}et\,B^*_N \Rightarrow \mathcal{B}et\,D$ becomes a quasi-axiom once more (i.e., $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,D \Rightarrow \mathcal{B}et\,D$). Also $\mathcal{B}et\,B^*_N \Rightarrow \mathcal{B}et\,C$ becomes $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,D \Rightarrow \mathcal{B}et\,C$ whose proof we obtained in the paragraph above. The third upper sequent now looks like $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,D \Rightarrow \mathcal{B}et\,B_N$ and it is proved as before. Proceeding downwards similarly as before we ultimately obtain a proof of $\mathcal{B}et\,\mathbb{B}_N \Rightarrow \mathcal{B}et\,D$. In analogous fashion we prove $\mathcal{B}et\,\mathbb{B}_N \Rightarrow \mathcal{B}et\,C$.

Finally, replace $\mathcal{B}et\,B^*_N$ with $\mathcal{B}et\,\mathbb{B}_N$ throughout the original $\mathcal{D}_1, \ldots \mathcal{D}_n$. The point is that the first and second upper sequents become $\mathcal{B}et\,\mathbb{B}_N \Rightarrow \mathcal{B}et\,C$ and $\mathcal{B}et\,\mathbb{B}_N \Rightarrow \mathcal{B}et\,D$ and we have already obtained proofs of these (the third upper sequent is handled similarly to before). Proceed downwards to obtain a proof of $\mathcal{B}et\,\mathbb{B}_N \Rightarrow B_i$ for every i. Every introduced cut was on B_N and hence the cut-rank of the final proof is $< \lceil \mathcal{B}et\,B_N \rceil$. □

The following lemma shifts the cut upward on the left premise of a cut when the right premise is principal, and uses Lemma 3.15 to reduce it.

Lemma 3.14 (permutation left) *Let $\mathcal{D}_l$ and $\mathcal{D}_r$ be* **HF** *proofs such that:*

(i) *$\mathcal{D}_l$ is a proof of $G \mid \Gamma_1, A^{\lambda_1} \Rightarrow \Delta_1 \mid \ldots \mid \Gamma_n, A^{\lambda_n} \Rightarrow \Delta_n$ and $\rho(\mathcal{D}_l) < \lceil A \rceil$;*

(ii) *A is a compound formula and $\mathcal{D}_r := H \mid \Sigma \Rightarrow A, \Pi$ ends with a right logical rule introducing the indicated occurrence of A, and $\rho(\mathcal{D}_r) < \lceil A \rceil$;*

Here each $\lambda_i > 0$. There is a transformation of these proofs into a **HF** *proof $\mathcal{D}$ of $G \mid H \mid \Gamma_1, \Sigma^{\lambda_1} \Rightarrow \Delta_1, \Pi^{\lambda_1} \mid \ldots \mid \Gamma_n, \Sigma^{\lambda_n} \Rightarrow \Delta_n, \Pi^{\lambda_n}$ with $\rho(\mathcal{D}) < \lceil A \rceil$.*

Proof. We distinguish cases according to the shape of A. If A is $\neg B$ or $B \rightarrow C$, the claim follows by Lemmas 3.11 and 3.12. If A is $\Box B$, $\bigcirc(B/C)$ or $\mathcal{B}et\, B$ the proof proceeds by induction on $|\mathcal{D}_l|$. If $\mathcal{D}_l$ ends in an initial sequent, then we are done. If $\mathcal{D}_l$ ends in a left rule introducing one of the indicated cut formulas, the claim follows by (i.h. and) Lemma 3.12. Otherwise, let (r) be the last inference rule applied in $\mathcal{D}_l$. The claim follows by the i.h., an application of (r) and/or weakening. Some care is needed to handle the cases in which r is $(s5')$, $(\Box R)$, $(\bigcirc R)$ or $(\mathcal{B}et)$ and A is not in the hypersequent context G. Notice that when $A = \Box B$ (resp. $A = \bigcirc(B/C)$) the conclusion of $\mathcal{D}_r$ is $\Sigma \Rightarrow \Box B, \Pi$ (resp. $\Sigma \Rightarrow \bigcirc(B/C), \Delta$), but we can safely use the "good"-shaped sequent $\Sigma^{\Box}, \Sigma^{O} \Rightarrow \Box B$ (resp. $\Sigma^{\Box}, \Sigma^{O} \Rightarrow \bigcirc(B/C)$), that allows cuts to be shifted upwards over all **HF** rules, and we apply weakening afterwards.

Let $A = \mathcal{B}et\, B$ and $\mathcal{D}_r$ ends in a $(\mathcal{B}et)$ rule with conclusion $\mathcal{B}et\, C, \Sigma \Rightarrow \mathcal{B}et\, B, \Pi$. If (r) is a $(\mathcal{B}et)$ rule introducing $\mathcal{B}et\, B$, the claim follows by Lemma 3.12. If (r) is (Bet_F), as in the proof below (to simplify the matter we omit both the internal and external contexts)

$$\dfrac{\{\mathcal{B}et\, \mathbb{D}, \mathcal{B}et\, B \Rightarrow D_j\}_{j=1,\ldots N} \qquad \begin{array}{c} \vdots\, d'_l \\ \mathcal{B}et\, \mathbb{D}, \mathcal{B}et\, B \Rightarrow B \end{array}}{\mathcal{B}et\, \mathbb{D}_i, \mathcal{B}et\, B \Rightarrow \mathcal{B}et\, D_i}\ (Bet_F)$$

we apply Lemma 3.13 to its premises (to get rid of the formula $\mathcal{B}et\, B$) and get

$$\{\mathcal{B}et\, \mathbb{D} \Rightarrow D_j\}_{j=1,\ldots N}.$$

The desired hypersequent $\mathcal{B}et\, \mathbb{D}_i, \mathcal{B}et\, C, \Sigma \Rightarrow \mathcal{B}et\, D_i, \Pi$ is simply obtained by applying the rule $(\mathcal{B}et_F)$ followed by (w). The case in which $\mathcal{D}_r$ ends in a $(\mathcal{B}et_F)$ rule is analogous. □

Lemma 3.15 (permutation right) *Let $\mathcal{D}_l$ and $\mathcal{D}_r$ be* **HF** *proofs where*

(i) *$\mathcal{D}_l$ concludes $G \mid \Gamma, A \Rightarrow \Delta$ and $\rho(\mathcal{D}_l) < \lceil A \rceil$*

(ii) *$\mathcal{D}_r$ concludes $H \mid \Sigma_1 \Rightarrow A^{\lambda_1}, \Pi'_1 \mid \ldots \mid \Sigma_n \Rightarrow A^{\lambda_n}, \Pi'_n$ with $\rho(\mathcal{D}_r) < \lceil A \rceil$.*

Here each $\lambda_i > 0$. There is a transformation of these proofs into a **HF** *proof $\mathcal{D}$ of $G \mid H \mid \Sigma_1, \Gamma^{\lambda_1} \Rightarrow \Pi'_1, \Delta^{\lambda_1} \mid \ldots \mid \Sigma_n, \Gamma^{\lambda_n} \Rightarrow \Pi'_n, \Delta^{\lambda_n}$ with $\rho(\mathcal{D}) < \lceil A \rceil$.*

Proof. Let (r) be the last inference rule applied in $\mathcal{D}_r$. If (r) is an axiom, then the claim holds trivially. If (one of) the indicated occurrence(s) of A is principal by (r) then the claim follows from Lemma 3.14. So suppose that no A is principal by (r). Proceed by induction on $|\mathcal{D}_r|$.

Consider the following analysis of (r): it acts only on H or is a rule other than $(s5')$, $(\Box R)$, $(\bigcirc R)$, $(\mathcal{B}et_F)$ and $(\mathcal{B}et)$; if it is $(\Box R)$, $(\bigcirc R)$, $(\mathcal{B}et_F)$ or

$(\mathcal{B}et)$ then the indicated A cannot be in the active premise component since that would make it principal; if (r) is $(s5')$ and A is in an active component of the conclusion it must be the component without any context restriction (it cannot be other since that should be empty). In all these cases the claim follows by applying the IH to the premise(s) followed by (r). □

Theorem 3.16 (Cut Elimination) *Cut elimination holds for* $\mathbf{HF} + cut$*.*

Proof. Define the *cut-multiset* $M_{\mathcal{D}}$ of $\mathcal{D}$ to be the multiset over the natural numbers $\mathbb{N}$ such that the multiplicity $M(n)$ of $n \in \mathbb{N}$ is the number of cut-rules in $\mathcal{D}$ with cut-rank n. We establish cut-elimination via induction on the Dershowitz-Manna [6] well-founded ordering over these multisets.

Let $\mathcal{D}$ be a $\mathbf{HF} + cut$ proof. Base case: $M_{\mathcal{D}} = \emptyset$ and hence $\mathcal{D}$ is cut-free. Inductive case: apply Lemma 3.15 to a subproof δ concluding a topmost cut in $\mathcal{D}$ (let the cut-formula be A). We thus obtain a new proof δ' whose cut-rank is $< \lceil A \rceil$. Let $\mathcal{D}'$ be the proof obtained from $\mathcal{D}$ by replacing δ with δ'. By inspection, $M_{\mathcal{D}'} <_m M_{\mathcal{D}}$ and hence the result follows by induction. □

Corollary 3.17 (Completeness) *Each theorem of* $\mathbf{F}$ *has a proof in* $\mathbf{HF}$*.*

4 A proof search oriented calculus for F

By modifying the calculus presented in Section 3, we obtain a decision procedure for the logic $\mathbf{F}$, and a complexity bound. The modified calculus $\mathbf{HF}^{+}$is based on the following ideas:

(i) Hypersequent component are considered as “set-based”: no duplication of formula is allowed within a component $\Gamma \Rightarrow \Delta$ of an hypersequent G.

(ii) In every rule the “principal” component(s) are kept in all premises, but not duplicated; thus hypersequents themselves are considered to be sets of components.

(iii) There are no redundant application of rules, in the sense that a rule is not applied (to a formula/component) if one of the premises of the rules is already contained in the conclusion.

(iv) There are no structural rules, except for the rule $(s5')$.

Restriction (i) is justified by the admissibility of internal contraction. As an example, by this restriction the backward application of $(\wedge L)$ will produce:

$$\frac{\Gamma, A, B \Rightarrow \Delta \,|\, G}{\Gamma, A, A \wedge B \Rightarrow \Delta \,|\, G} \quad \text{rather than} \quad \frac{\Gamma, A, A, B \Rightarrow \Delta \,|\, G}{\Gamma, A, A \wedge B \Rightarrow \Delta \,|\, G}$$

We display below the modified rules, we omit propositional rules; notice that the (O-L) rule does not need to be modified:

$$\frac{G \,|\, \Gamma^{\Box}, \Gamma^{O}, \Gamma' \Rightarrow \Pi' \,|\, G \,|\, \Gamma \Rightarrow \Delta}{G \,|\, \Gamma \Rightarrow \Delta \,|\, \Gamma' \Rightarrow \Pi'} \; (s5'_{new})$$

[6] $M <_m N$ iff $M \neq N$ and $M(k) > N(k)$ implies there is $k' > k$ such that $M(k') < N(k')$

$$\frac{\Gamma^{\Box,O}, A, \mathcal{B}et\,\neg A \Rightarrow B \mid G \mid \Gamma \Rightarrow \bigcirc(B/A), \Delta}{\Gamma \Rightarrow \bigcirc(B/A), \Delta \mid G}\,(\bigcirc R) \qquad \frac{\Gamma^{\Box,O}, B \Rightarrow A \mid G \mid \Gamma, \mathcal{B}et\,B \Rightarrow \Delta, \mathcal{B}et\,A}{\Gamma, \mathcal{B}et\,B \Rightarrow \Delta, \mathcal{B}et\,A \mid G}\,(\mathcal{B}et)$$

$$\frac{\Gamma^{\Box,O} \Rightarrow A \mid G \mid \Gamma \Rightarrow \Delta, \Box A}{\Gamma \Rightarrow \Delta, \Box A \mid G}\,(\Box R) \qquad \frac{\Gamma, A \Rightarrow \Delta \mid G \mid \Gamma, \Box A \Rightarrow \Delta}{\Gamma, \Box A \Rightarrow \Delta \mid G}\,(\Box L)$$

$$\frac{\{\Gamma^{\Box,O}, \mathcal{B}et\,\mathbb{D}, \mathcal{B}et\,B \Rightarrow D_i \mid G \mid S\}_{D_i \in \mathbb{D}} \qquad \Gamma^{\Box,O}, \mathcal{B}et\,\mathbb{D}, \mathcal{B}et\,B \Rightarrow B \mid G \mid S}{\Gamma, \mathcal{B}et\,\mathbb{D} \Rightarrow \mathcal{B}et\,B, \Delta \mid G}\,(\mathcal{B}et_F)$$

where $S = \Gamma, \mathcal{B}et\,\mathbb{D} \Rightarrow \mathcal{B}et\,B, \Delta$

It is tacitly assumed that contraction is applied in the premises (in particular for $(s5')$ rule), so that $\Box$ and O-formulas are not duplicated).

It is easy to see that the the calculus **HF**$^+$is sound and also complete, as a cut-free proof of **HF** can be simulated by **HF**$^+$ and *vice versa.*

Proposition 4.1 *Given an hypersequent G: $\vdash_{\mathbf{HF}} G$ iff $\vdash_{\mathbf{HF}^+} G$.*

Furthermore, observe that all rules are invertible, thus the order of application of rules within a derivation does not matter.

In order to obtain a decision procedure based on the calculus **HF**$^+$, we must avoid redundant application of rules in a backward proof search. First, let us define for two hypersequents G_1 and G_2 that $G_1 \sqsubseteq G_2$ if for every $\Gamma \Rightarrow \Delta \in G_1$ there is $\Gamma' \Rightarrow \Delta' \in G_2$ such that $\Gamma \subseteq \Gamma'$ and $\Delta \subseteq \Delta'$. We denote by $G_1 \sqsubset G_2$ the strict relation. Observe that for any rule R of **HF**$^+$:

$$\frac{G_1 \dots G_n}{G}\,(R)$$

we have $G \sqsubseteq G_i$ for $i = 1, \dots, n$. We say that an application of a rule R is *redundant* if for some $i \in \{1, \dots, n\}$, it holds $G_i \sqsubseteq G$. We say that a hypersequent G is *saturated* if it is not an axiom and all rule applications to it are redundant.

We adopt the following proof-search strategy: (i) no rule can be applied to an axiomatic sequent (ii) no redundant application of rule is allowed. The strategy preserves completeness.

Proposition 4.2 *Given an hypersequent G: if $\vdash_{\mathbf{HF}^+} G$ then G has a proof in* **HF**$^+$*according to the proof-strategy.*

From now on we restrict attention to derivations built according to the strategy. We show that any derivation with root sequent $\Rightarrow A$, for a formula A, is finite. To this purpose given a formula $A \in \mathcal{L}$, let $Sub(A)$ be the set of subformulas of A and $Sub^+(A) = Sub(A) \cup \{\mathcal{B}et\,\neg B : \bigcirc(C/B)$ occurs in $A\}$.

We now prove that the calculus **HF**$^+$provides a decision procedure for **F**.

Theorem 4.3 *Let $\mathcal{D}$ be a derivation in* **HF**$^+$ *with root $\Rightarrow A$ for a* **F***-formula A, then $\mathcal{D}$ is finite.*

Proof. Since the rules are analytic, given any hypersequent G occurring in $\mathcal{D}$, we have that for any $\Gamma \Rightarrow \Delta \in G$ we have $\Gamma \subseteq Sub^{+}(A)$ and $\Delta \subseteq Sub^{+}(A)$. But hypersequents are *sets* of components, thus it must be that for any $\Gamma \Rightarrow \Delta \in G$ and $\Gamma' \Rightarrow \Delta' \in G$ either $\Gamma \neq \Gamma'$ or $\Delta \neq \Delta'$. Thus G may have at most $2^{Sub^{+}(A)} \times 2^{Sub^{+}(A)}$ components, and each component has a size bounded by $Sub^{+}(A)$. Thus we can conclude that only finitely-many different hypersequents may occur in a derivation $\mathcal{D}$. By preventing repetitions of the same hypersequent on any branch (loop-checking), we get that every branch of $\mathcal{D}$ is finite. Since $\mathcal{D}$ is a finitely-branching tree, we can conclude that $\mathcal{D}$ is finite. □

Although the previous theorem ensures that any derivation is finite, it does not provide directly a decision algorithm for **F**.

Let n be the length of A as a string of symbols. Here is the decision procedure: we consider a non-deterministic algorithm which takes as input $\Rightarrow A$ and guesses a saturated hypersequent H: if it finds it, the algorithm answers "non-provable", otherwise, it answers "provable". By inspection, the size of the candidate saturated hypersequent H is $O(2^{2n})$. More concretely, the algorithm tries to build the candidate hypersequent H as follows: initialise a derivation with root $H_0 = \Rightarrow A$. Apply the rules backwards in an arbitrary but fixed order, choose non-deterministically a premise if there are more than one. In this way we generate a branch $\mathcal{B} = H_0, H_1, H_2 \ldots$. Observe that by the strategy, an application of a rule R to H_i is allowed only if H_i is not an axiom and that application of R is non-redundant, in this case it must be $H_i \sqsubset H_{i+1}$. The latter together with the observation that every hypersequent has size $O(2^{2n})$ implies that the length of every branch $\mathcal{B}$ is $O(2^{2n})$ and the last hypersequent H_k of $\mathcal{B}$ is either saturated or an axiom. Since every rule of $\mathbf{HF}^{+}$is invertible, unprovability of a hypersequent coincides with the existence of a branch rooted at that hypersequent whose leaf is saturated. Observe that all checks (whether H_i is an axiom, or is saturated, or whether an application of R to it is non-redundant) take at most quadratic time in the size of H_i.

The previous argument shows that non-provability in **F** can be decided in NEXP time. Whence we get:

Theorem 4.4 *Deciding if a formula is a theorem of* **F** *is in CoNEXP.*

Future work

The proposed calculus provides a preliminary complexity bound (CoNEXP) for theoremhood in **F**. Notice that CoNEXP is a worst-case bound, in practice there are several heuristics and techniques that could be adopted to reduce the complexity and get a more efficient proof system. Moreover, although the complexity of the decision problem was previously unknown, we expect that a better bound can be obtained by refining the rules of the calculus, in particular the $(\mathcal{B}et_F)$ rule which is the source of the exponential blow-up as in principle it has to be applied to any subset of $\mathcal{B}et$-formulas.

Furthermore we would like to investigate how to extract countermodels of non-valid formulas from failed derivations. This is a non-trivial task because of the limitedness condition that countermodels must satisfy.

Acknowledgements

X. Parent is supported by the Austrian Science Fund (FWF) through the ANCoR project [M-3240-N]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. R. Ramanayake acknowledges the support of the FWF project P33548. D. Rozplokhas is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 101034440.

References

[1] Åqvist, L., *Deontic logic*, in: D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic: Volume II*, Springer, Dordrecht, 1984 pp. 605–714.

[2] Avron, A., *A constructive analysis of RM*, J. of Symb. Logic **52** (1987), pp. 939–951.

[3] Avron, A., *The method of hypersequents in the proof theory of propositional non-classical logics*, in: *Logic: from foundations to applications*, OUP, New York, 1996 pp. 1–32.

[4] Belnap, N. D., Jr., *Display logic*, J. Philos. Logic **11** (1982), pp. 375–417.

[5] Chisholm, R., *Contrary-to-duty imperatives and deontic logic*, Analysis **24** (1963), pp. 33–36.

[6] Ciabattoni, A., N. Olivetti and X. Parent, *Dyadic obligations: Proofs and countermodels via hypersequents*, in: R. Aydoğan, N. Criado, J. Lang, V. Sanchez-Anguix and M. Serramia, editors, *PRIMA 2022* (2022), pp. 54–71.

[7] Danielsson, S., "Preference and Obligation," Filosofiska Färeningen, Uppsala, 1968.

[8] Fehige, C., "The Limit Assumption in Deontic (and Prohairetic) Logic," De Gruyter, Berlin, Boston, 1994 pp. 42–56.

[9] Giordano, L., V. Gliozzi, N. Olivetti and G. L. Pozzato, *Analytic tableaux calculi for KLM logics of nonmonotonic reasoning*, ACM Trans. Comput. Log. **10** (2009), pp. 18:1–18:47.

[10] Girlando, M., B. Lellmann, N. Olivetti and G. L. Pozzato, *Standard sequent calculi for Lewis' logics of counterfactuals*, in: *Proc. JELIA*, 2016, pp. 272–287.

[11] Goble, L., *Axioms for Hansson's dyadic deontic logics*, Filosofiska Notiser **6** (2019), pp. 13–61.

[12] Goré, R. and R. Ramanayake, *Valentini's cut-elimination for provability logic resolved*, in: *Advances in Modal Logic 7, papers from the seventh conference on "Advances in Modal Logic," held in Nancy, France, 9-12 September 2008*, 2008, pp. 67–86.

[13] Goré, R., R. Ramanayake and I. Shillito, *Cut-elimination for provability logic by terminating proof-search: Formalised and deconstructed using coq*, in: *Automated Reasoning with Analytic Tableaux and Related Methods - TABLEAUX 2021*, Lecture Notes in Computer Science **12842** (2021), pp. 299–313.

[14] Hansson, B., *An analysis of some deontic logics*, Noûs **3** (1969), pp. 373–398, reprinted in [15, pp. 121-147].

[15] Hilpinen, R., editor, "Deontic Logic," Reidel, Dordrecht, 1971.

[16] Kurokawa, H., *Hypersequent calculi for modal logics extending S4*, in: *New Frontiers in Artificial Intelligence*, LNCS **8417** (2013), pp. 51–68.

[17] Lehmann, D. and M. Magidor, *What does a conditional knowledge base entail?*, Artif. Intell. **55** (1992), pp. 1–60.

[18] Lewis, D., "Counterfactuals," Blackwell, Oxford, 1973.

[19] Minc, G., *Some calculi of modal logic*, Trudy Mat. Inst. Steklov **98** (1968), pp. 88–111.

[20] Parent, X., *A complete axiom set for Hansson's deontic logic DSDL2*, Log. J. IGPL **18** (2010), pp. 422–429.

[21] Parent, X., *Completeness of Åqvist's systems E and F*, Rev. Symb. Log. **8** (2015), pp. 164–177.

[22] Parent, X., *Preference semantics for Hansson-type dyadic deontic logic: a survey of results*, in: D. Gabbay, J. Horty, X. Parent, L. van der Torre and R. van der Meyden, editors, *Handbook of Deontic Logic and Normative Systems (vol. 2)*, College Publications, London, 2021 pp. 7–70.

[23] Sambin, G. and S. Valentini, *The modal logic of provability. The sequential approach*, J. Philos. Logic **11** (1982), pp. 311–342.

[24] Sawasaki, T. and K. Sano, *Term-sequence-dyadic deontic logic*, in: F. Liu, A. Marra, P. Portner and F. V. D. Putte, editors, *Deontic Logic and Normative Systems - 15th International Conference, DEON 2020/21* (2021), pp. 376–393.

[25] Valentini, S., *The modal logic of provability: cut-elimination*, J. Philos. Logic **12** (1983), pp. 471–476.

[26] van Benthem, J., P. Girard and O. Roy, *Everything else being equal: A modal logic for ceteris paribus preferences*, J. of Phil. Logic **38** (2009), pp. 83–125.

[27] Van De Putte, F. and S. C., *Preferential semantics using non-smooth preference relations*, J Philos Logic **43** (2014), pp. 903–942.

[28] van Fraassen, B., *The logic of conditional obligation*, J. of Phil. Logic **1** (1972), pp. 417–438.

5 Appendix

Proof. [Lemma 3.13 Base case $K = 0$]

Base case $K = 0$. There there are no $(\mathcal{B}et_F)$ rule instances containing $\mathcal{B}et\, B_N^*$. Define the *width of* $\mathcal{B}et\, B_N^*$ (terminology due to Valentini [25]) of a proof ending in a $(\mathcal{B}et_F)$ rule as the total number of upper sequents where $\mathcal{B}et\, B_N^*$ is introduced by a $(\mathcal{B}et)$ rule (this is the rule introducing $\mathcal{B}et$ in the antecedent, it should not to be confused with the $(\mathcal{B}et_F)$ rule!) with conclusion $\mathcal{B}et\, B_N^* \Rightarrow \mathcal{B}et\, C_w$ with $C_w \neq B_N$.

We establish the result by induction on the width W of the given proof which ends in a $(\mathcal{B}et_F)$ rule with premises $\mathcal{D}_1, \ldots, \mathcal{D}_n$. We proceed by case analysis on W.

Case $W = 0$. The upper sequents introduce $\mathcal{B}et\, B_N^*$ by weakening, or by a $(\mathcal{B}et)$ rule whose conclusion is $\mathcal{B}et\, B_N^* \Rightarrow \mathcal{B}et\, B_N$. The desired proof is obtained by replacing the occurrences of $\mathcal{B}et\, B_N^*$ in these upper sequents with $\mathcal{B}et\, \mathbb{B}_N$ as follows: the weakening on $\mathcal{B}et\, B_N^*$ is replaced with $\mathcal{B}et\, \mathbb{B}_N$, and the subproof ending in $\mathcal{B}et\, B_N^* \Rightarrow \mathcal{B}et\, B_N$ is replaced by a proof of $\mathcal{B}et\, \mathbb{B}_N \Rightarrow \mathcal{B}et\, B_N$ (itself obtained by applying $(\mathcal{B}et_F)$ to $\mathcal{D}_1, \ldots, \mathcal{D}_n$).

Case $W > 0$. Let $\mathbb{C} = \{C_1, \ldots, C_W\}$ be the set of upper sequents introducing $\mathcal{B}et\, B_N^*$ by a $(\mathcal{B}et)$ rule that conclude as $\mathcal{B}et\, B_N^* \Rightarrow \mathcal{B}et\, C_i$ with $C_i \neq B_N$.

Claim: If $\mathcal{B}et\, \mathbb{B}_N, \mathcal{B}et(\mathbb{C} \setminus S) \Rightarrow \mathcal{B}et\, C$ is provable with cut-rank $< \lceil \mathcal{B}et\, B_N \rceil$ for $S \subseteq \mathbb{C}$ and every $C \in S$, then the following premises of a $(\mathcal{B}et_F)$ rule are provable with cut-rank $< \lceil \mathcal{B}et\, B_N \rceil$:

$$\mathcal{B}et\, \mathbb{B}_N, \mathcal{B}et(\mathbb{C} \setminus S) \Rightarrow D \qquad (D \in \mathbb{B}_N \cup (\mathbb{C} \setminus S))$$

Proof of claim: let $S \subseteq \mathbb{C}$ be given. There are W occurrences of subproofs (spread across $\mathcal{D}_1, \ldots, \mathcal{D}_n$) that end in an upper sequent of the following form.

$$\frac{B_N \Rightarrow C}{\mathcal{B}et\, B_N^* \Rightarrow \mathcal{B}et\, C}\ (\mathcal{B}et) \text{ where } C \in \mathbb{C} \text{ and } C \neq B_N$$

For $C \in S$, replace the above with the following (the premise is the proof provided from the hypothesis).

$$\frac{\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et(\mathbb{C}\setminus S) \Rightarrow \mathcal{B}et\,C}{\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,B_N, \mathcal{B}et(\mathbb{C}\setminus S) \Rightarrow \mathcal{B}et\,C}\,(\mathrm{w})$$

For $C \in \mathbb{C}\backslash S$, replace instead with the 'obvious' proof (NB. $\mathcal{B}et\,C \in \mathcal{B}et(\mathbb{C}\backslash S)$)

$$\frac{\mathcal{B}et\,C \Rightarrow \mathcal{B}et\,C}{\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,B_N, \mathcal{B}et(\mathbb{C}\setminus S) \Rightarrow \mathcal{B}et\,C}\,(\mathrm{w})$$

In each of the W subproofs, $\mathcal{B}et\,B_N$ has been introduced by weakening. For this reason, proceeding downwards, we obtain the following premises of a $(\mathcal{B}et_F)$ rule with width 0 (the second row is obtained by a cut on $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,B_N, \mathcal{B}et(\mathbb{C}\setminus S) \Rightarrow B_N$ and $B_N \Rightarrow C$).

$$\begin{array}{ll} \mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,B_N, \mathcal{B}et(\mathbb{C}\setminus S) \Rightarrow B_i & \text{every } i \\ \mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,B_N, \mathcal{B}et(\mathbb{C}\setminus S) \Rightarrow C & C \in \mathbb{C}\setminus S \end{array}$$

Since the width is 0, we can remove the $\mathcal{B}et\,B_N$ from every sequent above (see Case $W = 0$) and hence the claim is proved.

Returning to the main proof (case $K = 0$), setting $S = \emptyset$, the hypothesis of the above claim is vacuously true and hence we obtain a proof of $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et\,\mathbb{C} \Rightarrow D$ for each $D \in \mathbb{B}_N \cup \mathbb{C}$ i.e. starting with the given proof which ends in a $(\mathcal{B}et_F)$ rule with premises $\mathcal{D}_1, \ldots, \mathcal{D}_n$, apply the transformation to every $C \in \mathbb{C}\setminus S (= \mathbb{C})$ that is described in the argument witnessing the claim.

Now apply $(\mathcal{B}et_F)$ to get $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et(\mathbb{C}\setminus\{C_1\}) \Rightarrow \mathcal{B}et\,C_1$. Applying the claim we get $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et(\mathbb{C}\backslash\{C_1\}) \Rightarrow D$ for each $D \in \mathbb{B}_N \cup (\mathbb{C}\backslash\{C_1\})$ and then from $(\mathcal{B}et_F)$ we get $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et(\mathbb{C}\setminus\{C_1, C_2\}) \Rightarrow \mathcal{B}et\,C_2$. We cannot apply the claim yet; we first need $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et(\mathbb{C}\backslash\{C_1, C_2\}) \Rightarrow \mathcal{B}et\,C_1$ and this is obtained in a similar manner. Apply the claim to get $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et(\mathbb{C}\setminus\{C_1, C_2\}) \Rightarrow D$ for each $D \in \mathbb{B}_N \cup (\mathbb{C}\setminus\{C_1, C_2\})$.

Now apply $(\mathcal{B}et_F)$ to get $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et(\mathbb{C}\setminus\{C_1, C_2, C_3\}) \Rightarrow \mathcal{B}et\,C_3$. Similarly obtain $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et(\mathbb{C}\setminus\{C_1, C_2, C_3\}) \Rightarrow \mathcal{B}et\,C_1$ and $\mathcal{B}et\,\mathbb{B}_N, \mathcal{B}et(\mathbb{C}\setminus\{C_1, C_2, C_3\}) \Rightarrow \mathcal{B}et\,C_2$, and then apply the claim. Proceeding in this manner we ultimately obtain the statement for $S := \mathbb{C}$ (i.e. $\mathcal{B}et\,\mathbb{B}_N \Rightarrow B_i$ for each i) so the lemma is proved.

□

Disambiguating permissions: A contribution from Mīmāṃsā

Agata Ciabattoni, Josephine Dik

TU Wien, Austria
agata@logic.at, josephine@logic.at

Elisa Freschi

University of Toronto, Canada
elisa.freschi@utoronto.ca

Abstract

The notion of permission has received less attention than obligation from the deontic logic community, that has often taken for granted the interdefinability of deontic operators (obligations, prohibitions and permissions). Yet, permission has proven to be a complex topic with various nuances that require careful treatment, and can lead to unwanted consequences if the interdefinability is kept. In contrast, the Sanskrit philosophical school of Mīmāṃsā refuted this interdefinability and instead established independent definitions for deontic concepts. This article focuses on the exploration of permission within Mīmāṃsā and its formalization through Hilbert axioms and semantics. We also compare the Mīmāṃsā approach to contemporary deontic logic discussions, and show that the central paradoxes of permission do not arise in the Mīmāṃsā paradigm.

Keywords: Permission; Interdefinability of deontic operators; Mīmāṃsā; Sanskrit philosophy; Deontic paradoxes; Free choice paradox; Neighbourhood semantics.

1 Introduction

Permission is of crucial importance in several settings, from law to ethics to artificial intelligence. Despite its significance, it has been the subject of fewer investigations in the deontic logic literature compared to obligation.

The concept of permission is inherently ambiguous and can be expressed in various manners such as “you are allowed to”, “ it is open for you to”, and “you have the right to”. Since the introduction of deontic logic by von Wright, permission has been often viewed simply as the dual of obligation [40], similar to how possibility serves as the dual of necessity in modal logic. Due to the unintuitive consequences (aka deontic paradoxes) mainly arising from this assumption, different varieties of permissions have been considered in the deontic logic literature; these include weak and strong permissions (e.g. [41,1,5,43,6]), bilateral and unilateral permissions (e.g. [12,26,27,23]), positive and negative permissions (e.g. [32,34]), and explicit, tacit or implicit permissions (e.g. [26]).

This paper contributes to the debate, by revealing and formalizing the concept of permission in Mīmāṃsā, which is one of the main Sanskrit philosophical schools and is a largely unexplored source for deontic investigations. Mainly active between the last centuries BCE and the 20th c. CE, Mīmāṃsā centred around the analysis of the prescriptive portions of the Vedas – the sacred texts of (what is now called) Hinduism. Mīmāṃsā authors interpreted the Vedas independently of the will of any speaker, as a consistent and self-sufficient corpus of laws. Thus, Mīmāṃsā authors have thoroughly discussed and analyzed normative statements in order to explain "what has to be done" in the presence of seemingly conflicting obligations. Since the Vedas are assumed to be not contradictory, Mīmāṃsā authors invested all their efforts in creating a consistent deontic[1] system. Key to their enterprise was the formulation of reasoning principles called *nyāya*s (see e.g. [19]), that lend themselves to a formalization through logic. Some *nyāya*s can be transformed into properties (Hilbert axioms) for the corresponding deontic operator in Mīmāṃsā, others (e.g. the specificity principle discussed in Kumārila's *Tantravārttika*) are instead metarules to resolve seeming contradictions in the Vedas.

The deontic theory of Mīmāṃsā has been progressively formalized through a series of works [14,29,8], each introducing new deontic operators and properties found in the original texts. The initial paper [14] presented the base logic, which considered only obligation, whose properties were extracted by analyzing around 40 *nyāya*s. Subsequently, prohibition was added in [29], and [8] included a weaker form of obligation, corresponding to elective duties, which are sacrifices to be performed only if one desires their specific outcome.

Our work has involved an interdisciplinary team effort that began with the discovery of the relevant *nyāya*s in Sanskrit texts, followed by their translation into English, their interpretation, and their formalization as Hilbert axioms. It is important to remark that our logics are solely based on principles extracted by Mīmāṃsā texts. Our aim is indeed to faithfully formalize the deontic theories of the Mīmāṃsā authors and use them to provide a better understanding of their debates, as well as new insights for contemporary deontic logic.

A distinctive feature of Mīmāṃsā deontics is the non-interdefinability of obligation and prohibition. As we have recently discovered, the independence of the deontic concepts extends also to permission, which is the focus of the present paper. Here we extend the logic discussed in [8] with the axioms for permission, and with newly formalised *nyāya*s, one of which corresponds to a version of the 'ought implies can' principle, see e.g. [9]. We propose a neighbourhood semantics for the resulting logic, which we call LM_P (Mīmāṃsā Logic with permission). To analyze the behaviour of LM_P we confront it with the best known deontic paradoxes concerning permission: free choice inference [42],

[1] Different Mīmāṃsā authors interpret commands differently (see [8]), but most of them looked at the Veda as a text having only deontic, i.e., normative authority.

Ross' paradox [37] and the paradox of the privacy act [22]. These paradoxes do not arise in LM_P; its well-behaved nature can be attributed to the millennia-old philosophical and juridical foundation upon which it is built.

The paper is organized as follows: Section 2 summarizes our previous findings on Mīmāṃsā deontics. Section 3 introduces the notion of permission in Mīmāṃsā and compares it to the literature of contemporary deontic logic. Mīmāṃsā permission is formalized in Section 4 by extending the logic in [8] with suitable Hilbert axioms and their semantics. In Section 5, the resulting logic is examined in light of the main deontic paradoxes related to permission, and it is demonstrated that it effectively addresses them.

Sanskrit sources: Throughout this paper, we refer to Jaimini's *Mīmāṃsā Sūtra* (or *Pūrva Mīmāṃsā Sūtra*, henceforth PMS, approximately 250 BCE) and Śabara's *Bhāṣya* 'commentary' thereon (henceforth ŚBh, approx. 5th c. CE), whose authority has been recognized by all Mīmāṃsā authors. We refer also to the following Mīmāṃsā texts: Kumārila's *Tantravārttika* (7th c., a key subcommentary on the PMS and ŚBh), and Rāmānujācārya's *Tantrarahasya* (14th c.?), as well as to a key text of Sanskrit jurisprudence (called Dharmaśāstra), Vijñāneśvara's *Mītākṣarā* (early 12th c.), a commentary on Yajñavalkya's code of norms.

2 Preliminaries on Mīmāṃsā Deontics

The Mīmāṃsā school focused on the rational interpretation and systematization of the prescriptive portions of the Vedas. These include *commands* of various kinds, such as prescriptions concerning the performance of sacrifices, and prohibitions applying either to the context of a sacrifice or to the entire life of a person (e.g. "One should not harm any living being"). Sometimes the commands seem to be contradictory, like in the case of the Śyena sacrifice, that should be performed if one wants to kill their enemy. [2] Mīmāṃsā thinkers introduced and applied metarules (called *nyāya*s) in order to rigorously analyze the Vedic commands and solve seeming contradictions among them.

The *nyāya*s are not listed explicitly by Mīmāṃsā authors, and have to be carefully distilled from their concrete applications within the texts. An example of a *nyāya* is "if a certain action is obligatory but it implies other activities, then these other activities are also obligatory" (Rāmānujācārya's *Tantrarahasya*).

Mīmāṃsā authors distinguish between obligations (*vidhi*) and prohibitions (*niṣedha*). [3] The former are determined by the fact of leading one to a desired result, if fulfilled, whereas the latter by the risk of sanction, if not fulfilled. This implies that negative obligations are different from prohibitions, and these two concepts are not mutually definable. "It is forbidden to lie" means that one will be liable to a sanction if one lies. "It is obligatory not to lie" means that one will receive a reward if one avoids any lie. Commands are always uttered with regard

[2] See [8] for the solutions to the Śyena controversy provided by the main Mīmāṃsā authors.

[3] Obligations and prohibitions in Mīmāṃsā have been discussed in [20], and formalized as suitable logics in [29,8].

to a specific person, called 'eligible' or 'responsible' (*adhikārin*), or to a specific situation in which an *adhikārin* might be in. In terms of deontic logic, this means that commands are always dyadic. For instance, the obligation to recite the Vedas is incumbent only on male members of the highest three classes who have undergone initiation, which can be rendered as $O(reciteVedas/initiated)$.

The use of logic to formalize Mīmāṃsā reasoning is justified by the rigorous theory of inference implemented by the school, that implicitly refers to logical principles and methods [14,19].

A further salient characteristic of Mīmāṃsā deontics is that commands have always one goal, hence they do not have conjunctions or disjunctions within them. A seemingly unitary command like "You should offer clarified butter and pour milk" would be interpreted as two separate commands, namely "You should offer clarified butter" and "You should pour milk".

Last, a metarule prescribes that commands should always convey something new (*apūrva*). A command that seems to prescribe an action one is already inclined to do should therefore be interpreted otherwise. For instance, "One should eat the five five-nailed animals" cannot be interpreted as enjoining the eating of certain animals, because one is naturally inclined to eat the meat of each animal. The command is instead interpreted as a prohibition of eating the meat of any other animal. A connected *nyāya* prescribes that the Vedas are always purposeful and do not enjoin anything without purpose. Altough the scope of these two *nyāya* may overlap, they are different as it is possible to imagine a norm being purposeful but not novel. As a consequence of these metarules, for instance prohibitions need to refer to actions one would be naturally inclined to undertake (*rāgaprāpta*) or that have already been enjoined (*śāstraprāpta*). Prohibiting something one would never undertake, e.g. "building an altar in the sky" would be purposeless and hence is not a viable interpretation of a command.

3 Permissions and new discoveries in Mīmāṃsā Deontics

One of the most striking features of Mīmāṃsā deontics is the non-interdefinability of deontic concepts, that also applies to the concept of permission. Its main characteristic is that a permission is always an exception to a prohibition or negative obligation. In Mīmāṃsā, saying "it is permitted to do X given Y", always entails that X is negatively obligatory or forbidden given a condition Z that is more general than Y. This can be illustrated by the following applications of an underlying *nyāya* (i.e. "A permission is always an exception to a pre-existing prohibition or negative obligation"):

(a) The permission to take a second wife can only occur as an exception to a general prohibition or negative obligation not to remarry (ŚBh on PMS 6.8.17–18).

(b) The permission to take up the occupation of a lower class in times of distress depends on the underlying prohibition to take up any occupation other than the ones admitted for one's own class (*Mitākṣarā* on 3.35).

(c) The permission to eat after buying Soma implies the prohibition to eat (or the obligation not to eat) before it (*Tantravārttika* on 1.3.4).
(d) The permission to sell while being a Brāhmaṇa in distress implies the prohibition to sell while being a Brāhmaṇa in normal circumstances (*Mitākṣarā* on 3.35).

Thus, these permissions are interpreted as presupposing an underlying prohibition or negative obligation, and not as stand-alone permissions.

The permission to sell while being a Brāhmaṇa in distress, for instance, implies that a Brāhmaṇa not in distress should not be selling anything. Similarly, the permission to take up the occupation of a lower class in times of distress depends on the underlying prohibition to take up any occupation other than the ones admitted for one's own class (see Vijñāneśvara's *Mitākṣarā* commentary on Yājñavalkya 3.35) and the permission to eat after a certain moment of the sacrifice implies the prohibition to eat before it (*Tantravārttika* on 1.3.4).

Hence permissions only make sense for Mīmāṃsā authors with regard to acts which were previously prohibited or the abstention from which was obligatory. To define the realm of "whatever is not prohibited is permitted", Mīmāṃsā authors introduce the concept of "normatively indifferent actions". These are actions that are possible, but neither prohibited *nor* enjoined (nor permitted in the Mīmāṃsā sense) and that constitute most of our everyday life. Normatively indifferent actions are the ones on which normative texts make an intervention. For instance, offering a ritual substance is not permitted in a Mīmāṃsā sense, because it is enjoined. In the following, we will call whatever is neither prohibited nor permitted nor enjoined "extra-normative". In sum, for Mīmāṃsā there are either normed actions (enjoined, prohibited or permitted) or extra-normative ones.

A last feature of Mīmāṃsā permission is the following: if X is permitted given Y, doing X is not on the same level as not doing it, or as doing X while X is an extra-normative action. Rather, permissions allow an option that is less desirable than its counterpart. One of the main consequences of this approach is that performing a permitted X exposes one to the risk of restrictions, insofar as the permitted action is actually an action one should have "better-not" performed. Thus, eating after having bought Soma is permissible, but not eating is the preferred option (for more details, see [18], Section 4).

Related to permission, we have newly (identified and) formalized a characteristic of Mīmāṃsā deontics, that is a version of the 'ought implies can' principle, usually attributed to Immanuel Kant, see [38], and that in Mīmāṃsā's case can be formulated as "each command must be actionable", thus including the claim that also forbidden entails can. This metarule is extracted from the *nyāya*s "Prescriptions can only prescribe actions that can be performed" and "Prohibitions can only prohibit actions that can be performed", whose application is found below:

(e) Commands prescribing complicated sacrifices in order to get *svarga* (that is, heaven, to be understood as happiness) are addressed to men who are able to perform them (see *Tantravārttika* on 1.3.4).

(f) The seeming prohibition “The fire is not to be kindled on the earth, nor in the sky, nor in heaven” cannot be taken as a prohibition, because fire cannot be kindled in the sky nor in heaven (see ŚBh on 1.2.5 and 1.2.18).

The metarule regarding novelty (*apūrva*, see Section 2) also implies that it is impossible to have more than one deontic operator applied to the same action under the same circumstances.

Rather, each deontic operator needs to make a novel intervention and is therefore applied to an extra-normative situation, or, in the case of permissions, to a pre-existing negative obligation or prohibition. With regard to permissions, this also means that the same action cannot be at the same time obligatory and permitted given the same circumstances (pace SDL [40]), since the operator for permission would not add anything novel if applied to a situation already normed by the deontic operator for obligation. For instance, if one already knows that male married Brahmins ought to perform a certain ritual at dawn, receiving the information that it is permitted to perform the same ritual at the same time and given the same circumstances would be redundant and purposeless, and no command in the Veda can be purposeless.

3.1 Mīmāṃsā Permission vs Permission in Deontic Logic

The interdefinability between obligation and permission is an old problem in Deontic Logic, dating back to the observation by Von Wright in [40] of the similarity with the relation between necessity and possibility. The deontic axiom D included in Standard Deontic Logic SDL therein introduced says that obligation implies permission.

As emphasized in [1], a main problem with this interdefinability is that the resulting system does not allow for gaps. If everything that is not permitted is prohibited and everything that is not prohibited is permitted, then any normative system would regulate all possible states of affairs. This is counterintuitive since not all situations are subject to regulation, as also acknowledged by the Mīmāṃsā school and its recognition of extra-normative actions.

Mīmāṃsā’s concept of extra-normativity aligns with the idea of indifference as defined by McNamara in relation to supererogation [33]. In McNamara’s definition, an indifferent action is neither obligatory nor forbidden. Moreover, the author links an operator for indifference with one for “moral significance” and thus differentiates between indifference and supererogation. Both indifferent and supererogatory actions are neither obligatory nor forbidden, but supererogatory actions hold moral significance.

In [41], von Wright treats the notion of permission more carefully than in his previous writings and introduces a distinction between weak permission and strong permission. Weak permission is permission as the absence of prohibition, whereas strong permission is a modality by itself. The latter is defined as follows: (i) an act “will be said to be permitted in the strong sense if it is not forbidden but subject to norm”, and (ii) “an act is permitted in the strong sense if the authority has considered its normative status and decided to permit it". A formalization of strong permission is contained, e.g., in [36]. Like in the case of Mīmāṃsā, it functions as a dyadic operator, but, unlike in Mīmāṃsā, it can be granted under general conditions (and not just as an exception to a prohibition or negative obligation), and all tautologies are trivially permitted, which is not the case in Mīmāṃsā.

Many authors have attempted a formalization of von Wright's definition of strong permission, mainly with the purpose of obtaining a consistent formalization of the so-called 'free choice inference', introduced in [42]. This inference is of the form 'If it is permitted to do A or B, it is permitted to do A and it is permitted to do B'. Generally, a disjunction of permissions implies that any of the disjuncts is a possible option, and this is clearly an inference scheme that is desirable for a permission to follow. However, accepting the free choice inference might lead to deriving counterintuitive conclusions, e.g., an obligation to pay your taxes implies a permission to murder someone. Among the works that have endeavored to establish a formalization of free choice permissions that are immune to undesired consequences are [3,5,6,16]. The introduced systems are quite complex, and use, e.g., substructural logics as underlying logics or semantical elements added to the language.

Hansson's paper [26] explores a third form of permission: implicit permission, which is implied by an obligation. For instance, the obligation to testify in court implies the permission to enter the courtroom. In contrast, for Mīmāṃsā an act cannot be both obligatory and permitted under the same circumstance and the obligation to perform X extends to the obligation to perform whatever is necessarily entailed by X. Thus, entering the courtroom is not the content of an implicit permission but of an obligation.

Alchourrón famously recounts a story (originally from [17]) about a hunting tribe and its new chief, who emits a norm permitting hunting on certain days, but without prohibiting it on the others. The tribe is utterly dissatisfied, because one expected from the chief an intervention in the status quo ("The moral of this story is valuable. It shows that purely permissory norms are of little if any practical interest" [2]). Alchourrón's conclusion, is different from the Mīmāṃsā one, as he highlights the importance of permissions in the case of more than one source of norms, see [2]. However the tribe reasoned according to Mīmāṃsā principles, based on which each command needs to change something which was previously the case (see the novelty requirement discussed above and the examples mentioned in Section 3).

A Mīmāṃsā permission is always an exception to a more general prohibition or negative obligation. This approach reflects a common practice in normative texts, such as legal codes in European jurisprudence, where permissions are typically stated only when there is an expectation of the opposite due to a general prohibition. Norms granting permissions usually derogate from what is stated in other norms, as Bouvier notes in his definition of permission in his legal dictionary in [11]. He distinguishes between express permissions that "derogate from something which before was forbidden," and implied permissions, "which arise from the fact that the law has not forbidden the act to be done". The latter are therefore different from Hansson's "implicit permission" and rather correspond to what Hansson calls "tacit permissions" in [26], and to what von Wright calls "weak permissions" in [41]. Similarly, the idea that permissions grant one a different degree of freedom if compared to the non-normed space of indifferent actions is neatly reflected by the comparison of cases like "You are permitted to run 2km per day" (said by a physician to her patient, who is recovering from a heart attack), as opposed to the same person's freedom to run prior to the heart attack. The permission rules the realm of running by introducing a space of possibility that is, however, not as absolute as the space of extra-normative actions. Accordingly, permitted actions are actions one would be naturally inclined to do, prior to the intervention of a normative text prohibiting them (or obliging one to refrain from them). In Mīmāṃsā deontics, it would not make sense to have a permission that regards impossible actions like flying or undesirable actions like harming oneself (assuming that harming oneself is not desirable for anyone). The Mīmāṃsā position is neatly distinguished from the one of, e.g., [26], who thinks that introducing permissions even in the absence of general prohibitions are useful to define rights.

A last trait of Mīmāṃsā permissions is that they always lead to less desirable options. This offers a solution to seeming problems like the "Interrupted promise", discussed by Zylberman [44]. There, one promises to participate in a conference, but then one's daughter has an accident and the previous duty is overruled by the duty to stand by one's daughter during surgery. Zylberman notes that despite having permission to withdraw, there is still an obligation to apologize or make reparations to the conference organizers. This sentiment contradicts the standard account of permissions, which does not mandate such actions. For instance, if it is permitted to drive at 18, no 18-years-old is expected to apologise because they are in fact driving. By contrast, the "interrupted promise" problem is instantly solved if we realise that the permission Zylberman is referring to is a Mīmāṃsā permission ("better-not") and therefore requires some expiation (such as offering an apology).

The concept of preference in deontic logic is well known, see, e.g., [15,24,39,30,4,25]. However, its application to a "less preferred" permission has not been explored in depth. We defer to future research the examination of the preference aspect of permission. Instead, we focus below on the formalization of the remaining properties.

4 Formalizing Mīmāṃsā Permission

Following a bottom-up approach of extracting deontic principles from the Mīmāṃsā texts, we transform the properties of the permission operator into suitable Hilbert axioms, which are added to the logic LKu^+ of [8]. We call the resulting logic LM_P (Mīmāṃsā Logic with permission). In this section, we present and justify its Hilbert axiomatization, we introduce a neighbourhood semantics, and demonstrate soundness, completeness and consistency of LM_P.

Ax1. $(\boxed{\mathsf{U}}(\phi \to \psi) \wedge \mathcal{O}(\phi/\theta) \wedge \neg\boxed{\mathsf{U}}\psi) \to \mathcal{O}(\psi/\theta)$
Ax2. $(\boxed{\mathsf{U}}(\phi \to \psi) \wedge \mathcal{F}(\psi/\theta) \wedge \Diamond_{\mathsf{U}}\phi) \to \mathcal{F}(\phi/\theta)$
Ax3. $\neg(X(\phi/\theta) \wedge X(\neg\phi/\theta))$ for $X \in \{\mathcal{O}, \mathcal{F}\}$
Ax4. $\neg(\mathcal{O}(\phi/\theta) \wedge \mathcal{F}(\phi/\theta))$
Ax5. $(\boxed{\mathsf{U}}(\psi \leftrightarrow \theta) \wedge X(\phi/\psi)) \to X(\phi/\theta)$ for $X \in \{\mathcal{O}, \mathcal{F}\}$
Ax6. $(\Diamond_{\mathsf{U}}(\phi \wedge \theta) \wedge \mathcal{O}(\phi/\top) \wedge \mathcal{O}(\theta/\top)) \to \mathcal{O}(\phi \wedge \theta/\top)$

Table 1
Axioms regarding obligation and prohibition from [8]

4.1 Syntax

The logic LM_P is an extension of the logic LKu^+ of [8]. Recall that the language of LKu^+ consists of the modalities[4] $\mathcal{O}(\phi/\psi)$ and $\mathcal{F}(\phi/\psi)$ for obligation and prohibition (read as "ϕ is obligatory/prohibited given ψ"). Here we add the permission operator $\mathcal{P}(\phi/\psi)$, to be read as "ϕ is permitted, given ψ". This operator is treated as a primitive modality, that is, $\mathcal{P}(\phi/\psi)$ is not equivalent to $\neg\,\mathcal{F}(\phi/\psi)$ or $\neg\,\mathcal{O}(\neg\phi/\psi)$. As explained in Section 2, all the deontic operators in Mīmāṃsā are dyadic. The language $\mathcal{L}_{LM_P}$ is defined as follows:

$$\phi ::= p \in Atom \mid \neg\phi \mid \phi \vee \phi \mid \boxed{\mathsf{U}}\phi \mid \mathcal{O}(\phi/\phi) \mid \mathcal{F}(\phi/\phi) \mid \mathcal{P}(\phi/\phi)$$

(where *Atom* is the set of atomic propositions). We take the classical logic[5]. connective $\neg$ and $\vee$ as primitive, and define $\wedge$, $\to$, $\leftrightarrow$ in the usual way. The constants $\top$ and $\bot$ are abbreviations for $\neg\phi \vee \phi$ and $\neg\top$, respectively. $\boxed{\mathsf{U}}$ is the universal S5 modality, read as 'in all scenarios, ϕ is true' and its dual $\Diamond_{\mathsf{U}}\phi = \neg\boxed{\mathsf{U}}\neg\phi$ as 'there is at least one scenario where ϕ is true'.

Definition 4.1 The logic LM_P extends LKu^+ – whose axiomatization consists of the axiomatization for the modal logic S5 for $\boxed{\mathsf{U}}$ and the axioms of Table 1 – with the following axioms:

P1. $\mathcal{P}(\phi/\psi) \to (\mathcal{F}(\phi/\top) \vee \mathcal{O}(\neg\phi/\top))$
P2. a) $\neg(\mathcal{P}(\phi/\psi) \wedge \mathcal{F}(\phi/\psi))$
b) $\neg(\mathcal{P}(\phi/\psi) \wedge \mathcal{O}(\phi/\psi))$
c) $\neg(\mathcal{P}(\phi/\psi) \wedge \mathcal{O}(\neg\phi/\psi))$
P3. $(\mathcal{O}(\phi/\psi) \vee \mathcal{F}(\phi/\psi)) \to \Diamond_{\mathsf{U}}(\phi \wedge \psi) \wedge \neg\boxed{\mathsf{U}}\phi$
P4. a) $(\boxed{\mathsf{U}}(\psi \leftrightarrow \theta) \wedge \mathcal{P}(\phi/\psi)) \to \mathcal{P}(\phi/\theta)$
b) $(\boxed{\mathsf{U}}(\phi \leftrightarrow \psi) \wedge \mathcal{P}(\phi/\theta)) \to \mathcal{P}(\psi/\theta)$
P5. $(\mathcal{P}(\phi/\psi) \wedge (\mathcal{F}(\phi/\theta) \vee \mathcal{O}(\neg\phi/\theta))) \to \boxed{\mathsf{U}}(\psi \to \theta)$

[4] The logic LKu^+ formalizes the deontic theories of two main Mīmāṃsā authors: Kumārila and Prabhākara (both 7 CE?). Their theories differ from the way elective duties are interpreted: as an obligation for Prabhākara, and as a recipe that guarantees to obtain a desired result, for Kumārila. The latter has been formalized in [8] with a modality $\mathcal{E}(\phi/\psi)$ having no deontic force. As this modality does not interact with the deontic operators, to simplify the matter we omit it from the language of LM_P.

[5] The classical logic base is justified by the presence, e.g., of the reduction ad absurdum law in Mīmāṃsā, see [14]

Introduced in [8] the axioms in Table 1 are based on the following principles extracted from suitable *nyāya*s:

1. If the accomplishment of an action presupposes the accomplishment of another action, the obligation to perform the first prescribes also the second. Conversely, if an action necessarily implies a prohibited action, this will also be prohibited. This corresponds to the *nyāya* given as an example in Section 2, and formalized by Ax1 and Ax2.
2. Two actions that exclude each other can neither be prescribed nor prohibited simultaneously to the same group of eligible people under the same conditions. This principle is the base for Ax3 and Ax4.
3. If two sets of conditions always identify the same group of eligible agents, then a command valid under the conditions in one of the sets is also enforceable under the conditions in the other set. This is formalized by Ax5.
4. If two fixed duties are prescribed and compatible, their conjunction is obligatory as well. This corresponds to Ax6.

Remark 4.2 In this paper we use a slightly different formulation of the axioms Ax1 and Ax2, w.r.t. [8], as their original version leads to contradictions in the presence of our new axiom P3. Ax1 was indeed presented as $(\boxed{\scriptstyle U}(\phi \to \psi) \wedge \mathcal{O}(\phi/\theta)) \to \mathcal{O}(\psi/\theta)$. Since $\boxed{\scriptstyle U}(\phi \to \top)$ is true for any formula ϕ, we would derive $\mathcal{O}(\top/\theta)$ whenever we have $\mathcal{O}(\phi/\theta)$ for any ϕ and θ, contradicting axiom P3. Ax2 was presented in [8] as $(\boxed{\scriptstyle U}(\phi \to \psi) \wedge \mathcal{F}(\psi/\theta)) \to \mathcal{F}(\phi/\theta)$. The formula $\boxed{\scriptstyle U}(\bot \to \psi)$ is true for any formula ψ, and therefore we derive $\mathcal{F}(\bot/\theta)$ from $\mathcal{F}(\psi/\theta)$ for any ψ and θ, contradicting P3, as well.

We discuss the properties of permission that led to the definition of axioms P1-P5 in Def. 4.1. We start by presenting the abstract principles behind them.

(i) Permissions are always exceptions to more general prohibitions or negative obligations.

This principle is extracted from the *nyāya* applied in (a)-(d) from Section 3, and is the base for axioms P1 and P5. P1 represents the fact that a permission is always an exception to a general prohibition or negative obligation (cf. (a)-(c)). From the application (d), we conclude that if something is allowed in one context and prohibited (or negatively obliged) in another, the context of the prohibition or negative obligation is more general, as formalized by axiom P5.

(ii) No more than one deontic operator can be applied to the same action under the same circumstances.

In the domain of Mīmāṃsā deontics, this principle represents a foundational metarule (cf. the *apūrva*-metarule discussed in Section 2 and 3) and justifies P2a-P2c. These axioms are similar to Ax4, but extended to permission. Especially interesting is axiom P2b, which states that an action cannot be permitted as well as obligatory under the same circumstances, contradicting the often-accepted inference in deontic logic that obligation implies permission.

(iii) Commands entail possibility.

The formalization of this principle is accomplished through Axiom P3, which does not pertain to permission. The principle has been extracted from various contexts, summarized by the *nyāya*-applications (e), corresponding to 'ought implies can', and (f), corresponding to 'forbidden implies can' (cf. Section 3). As commands must be meaningful, this axiom also excludes the possibility of obligatory or prohibited tautologies. Although we have not found an explicit statement that principle (iii) applies also to permissions, the fact that permitted actions are exceptions to prohibited or negatively obliged (possible) actions, is enough to conclude that this axiom should be present; as shown by Lemma 4.4 it is indeed derivable in LM_P.

Axiom P4a and P4b do not follow from any explicit discussion by Mīmāṃsā authors. P4a is implicitly used in Dharmaśāstra discussions of permissions under extreme circumstances. For instance, Vijñāneśvara states that it is permitted to sell certain vegetables if one has assumed the occupation of the *vaiśya* class, and then refers to the permission to sell the same vegetables if one is working as a merchant, given that assuming the occupation of a *vaiśya* implies being a merchant (*Mitākṣarā* on Yājñavalkya 3.35). Axiom P4b is also implicitly used in the same context when $\mathcal{P}$(act as a *vaiśya*/(being a Brāhmaṇa $\wedge$ being in distress)) leads to the $\mathcal{P}$(selling/(being a Brāhmaṇa $\wedge$ being in distress)) because acting as a *vaiśya* is synonymous of selling.

Remark 4.3 In contrast with obligation and prohibition (as well as contrary to the notion of permission in [36]), LM_P does not contain a monotonicity axiom for permission, i.e., $(\mathcal{P}(\phi/\theta) \wedge \boxed{\scriptstyle U}(\phi \to \psi)) \to \mathcal{P}(\psi/\theta)$. The main reason is that we have not found it in Mīmāṃsā texts. It is also unlikely to find it as this axiom would lead to unwanted consequences. For instance, from "eating meat implies being alive" and $\mathcal{P}$(eating meat/during extreme circumstances), follows $\mathcal{P}$(being alive/during extreme circumstances) which is not meaningful as we have no control over being alive. Additionally, as shown by the following derivation, monotonicity of permissions would imply an unconditional prohibition or negative obligation for any other feasible action:

(i) $\mathcal{P}(\phi/\theta) \to \mathcal{P}(\phi \vee \psi/\theta)$ (monotonicity for permissions)

(ii) $\mathcal{P}(\phi \vee \psi/\theta) \to (\mathcal{F}(\phi \vee \psi/\top) \vee \mathcal{O}(\neg(\phi \vee \psi)/\top))$ (P1)

(iii) $\boxed{\scriptstyle U}(\psi \to (\phi \vee \psi)) \wedge \mathcal{F}(\phi \vee \psi/\top) \wedge \Diamond_{U}\psi \to \mathcal{F}(\psi/\top)$ (Ax2)

(iv) $\boxed{\scriptstyle U}((\neg\phi \wedge \neg\psi) \to \neg\psi) \wedge \mathcal{O}(\neg(\phi \vee \psi)/\top) \wedge \Diamond_{U}\psi \to \mathcal{O}(\neg\psi/\top)$ (Ax1)

(v) $\mathcal{P}(\phi/\theta) \wedge \Diamond_{U}\psi \to (\mathcal{F}(\psi/\top) \vee \mathcal{O}(\neg\psi/\top))$ (from (i)-(iv))

Lemma 4.4 *The following formulas are derivable in LM_P:*

1. $\boxed{\scriptstyle U}(\phi \to \psi) \to \neg(\mathcal{O}(\phi/\theta) \wedge \mathcal{P}(\psi/\theta))$
2. $\boxed{\scriptstyle U}(\phi \to \psi) \to \neg(\mathcal{F}(\psi/\theta) \wedge \mathcal{P}(\phi/\theta))$
3. $\mathcal{P}(\phi/\psi) \to \Diamond_{U}\phi \wedge \neg\boxed{\scriptstyle U}\phi$
4. $\neg(\mathcal{P}(\phi/\theta) \wedge \mathcal{P}(\neg\phi/\theta))$

Proof. 1. follows by Ax1 and P2b. 2. follows by Ax2 and P2a. 3. follows from by axiom P1 and P3. 4. follows from P1, Ax3 and Ax4. □

The first two formulas from Lem. 4.4 can be viewed as generalizations of the D-axiom for permission. Formula 3, that will be utilized in our formalization of the free choice inference in the next section, constitutes a variation of the 'commands entail possibility' principle for permissions. Although formula 4 is not a property of permission in natural language, in the context of Mīmāṃsā, permissions are treated as exceptions to general prohibitions or negative obligations and there cannot be a prohibition or negative obligation regarding both a particular action and its negation.

4.2 Semantics

In line with [8], we use neighbourhood semantics to model LM_P.

Neighbourhood semantics generalizes Kripke semantics. It consists of a set of worlds W and a valuation function V, and contains neighbourhood functions N_x that map a world to a set of ordered pairs of sets of worlds. Each of the three modalities, obligation, permission and prohibition, has its own neighbourhood function. For example, let $w \in W$, if (X, Y) is in w's obligation-neighbourhood, this means that X represents the worlds of compliance 'from the point of view' of Y. Then, if X is exactly the set of worlds where ϕ is true, and Y is exactly the set of worlds where ψ is true, then $\mathcal{O}(\phi/\psi)$ is true in w.

Definition 4.5 An LM_P-frame $F = \langle W, N_{\mathcal{O}}, N_{\mathcal{P}}, N_{\mathcal{F}} \rangle$ is a tuple where $W \neq \emptyset$ is a set of worlds $w, v, u, \ldots$ and $N_{\chi} : W \mapsto P(P(W) \times P(W))$ is a neighbourhood function for $\chi \in \{\mathcal{O}, \mathcal{P}, \mathcal{F}\}$. Let $X, Y, Z \subseteq W$, F satisfies the following:

(i) If $(X, Z) \in N_{\mathcal{P}}(w)$ then $(X, W) \in N_{\mathcal{F}}(w)$ or $(\overline{X}, W) \in N_{\mathcal{O}}(w)$.
(ii) If $(X, Z) \in N_{\mathcal{P}}(w)$ then $(X, Z) \notin N_{\mathcal{F}}(w)$ and $(X, Z) \notin N_{\mathcal{O}}(w)$.
(iii) If $(X, Z) \in N_{\chi}(w)$ then $X \bigcap Z \neq \emptyset$ and $X \neq W$ for $(\chi \in \{\mathcal{O}, \mathcal{F}\})$.
(iv) If $(X, Y) \in N_{\mathcal{P}}(w)$ and $(X, Z) \in N_{\mathcal{F}}(w)$ or $(\overline{X}, Z) \in N_{\mathcal{O}}(w)$ then $Y \subset Z$.
(v) If $(X, Z) \in N_{\mathcal{P}}(w)$ then $(\overline{X}, Z) \notin N_{\mathcal{O}}(w)$.
(vi) If $(X, Z) \in \mathcal{N}_{\mathcal{O}}(w)$ and $X \subseteq Y$ and $Y \neq W$, then $(Y, Z) \in \mathcal{N}_{\mathcal{O}}(w)$.
(vii) If $(X, Z) \in \mathcal{N}_{\mathcal{F}}(w)$ and $Y \subseteq X$ and $Y \neq \emptyset$, then $(Y, Z) \in \mathcal{N}_{\mathcal{F}}(w)$.
(viii) If $(X, Y) \in \mathcal{N}_{\mathcal{X}}(w)$, then $(\overline{X}, Y) \notin \mathcal{N}_{\mathcal{X}}(w)$ for $\mathcal{X} \in \{\mathcal{O}, \mathcal{F}\}$.
(ix) If $(X, Z) \in \mathcal{N}_{\mathcal{O}}(w)$ then $(X, Z) \notin \mathcal{N}_{\mathcal{F}}(w)$.
(x) If $X \bigcap Y \neq \emptyset$ and $(X, W), (Y, W) \in \mathcal{N}_{\mathcal{O}}(w)$, then $(X \bigcap Y, W) \in \mathcal{N}_{\mathcal{O}}(w)$.

An LM_P-model is a tuple $M = \langle F, V \rangle$ where F is an LM_P-frame and V is a valuation function mapping atomic propositions from *Atom* to sets of worlds.

Note that (i) corresponds to axiom P1, (ii) and (vi) to axioms P2a-c, (iii) to axiom P3 and (iv) to P5. Moreover, (vi) and (vii) correspond to axioms Ax1 and Ax2, expressing the property of monotonicity in the first argument of the deontic operators; these conditions are based on the ones in [8], adjusted to comply with our new version of the monotonicity axioms (see Remark 4.2). (viii) corresponds to Ax3, avoiding the accumulation of deontic operators, (ix) corresponds to Ax4, and (x) to Ax6. Axioms P4a, P4b and Ax5 hold in any neighbourhood model [13] and do not require explicit conditions.

Definition 4.6 Let M be a LM_P-model and $\|\phi\| = \{w \in W \mid M, w \vDash \phi\}$. We define the satisfaction of a formula $\phi \in \mathcal{L}_{LM_P}$ at any $w \in W$ as follows:

$$\begin{array}{lll}
M, w \vDash p & \text{iff} & w \in V(p), \text{ for } p \in Atom \\
M, w \vDash \neg\phi & \text{iff} & M, w \nvDash \phi \\
M, w \vDash \phi \vee \psi & \text{iff} & M, w \vDash \phi \text{ or } M, w \vDash \psi \\
M, w \vDash \boxed{\mathsf{U}}\,\phi & \text{iff} & \text{for all } w_i \in W \;\; M, w_i \vDash \phi \\
M, w \vDash \Diamond_{\mathsf{U}}\,\phi & \text{iff} & \text{there exists a } w_i \in W \;\; M, w_i \vDash \phi \\
M, w \vDash \mathcal{X}(\phi/\psi) & \text{iff} & (\|\phi\|, \|\psi\|) \in \mathcal{N}_{\mathcal{X}}(w) \text{ for } \mathcal{X} \in \{\mathcal{O}, \mathcal{F}, \mathcal{P}\}
\end{array}$$

We say a formula ϕ *holds* in a model M iff $M, w \vDash \phi$ for each $w \in W$.

Using the strategy outlined in [10] and their corresponding definitions, we demonstrate that the axioms are sound and complete relative to the given neighbourhood semantics.

Definition 4.7 A formula ϕ is *valid* in LM_P, if for all worlds w in all LM_P-models M it is the case that $M, w \vDash \phi$.

A formula ϕ is a *theorem* of LM_P, if it is derivable using only the axioms of LM_P, modus ponens and necessitation rule for $\boxed{\mathsf{U}}$.

Theorem 4.8 (Soundness) *If a formula ϕ is a theorem of LM_P, then ϕ is valid.*

Proof. We show that all axioms of LM_P are true in all worlds of any LM_P-model M. For each axiom, we assume that the antecedent holds in a world, and use the neighbourhood restrictions and the truth conditions of Def. 4.6 to derive the intended consequent. Showing that modus ponens and the necessitation rule for $\boxed{\mathsf{U}}$ preserve validity is easy. We only detail the case of axiom P1 – the main property of Mīmāṃsā permission– as all other axioms are proven similarly. Assume that $\mathcal{P}(\phi/\psi) \to (\mathcal{F}(\phi/\top) \vee \mathcal{O}(\neg\phi/\top))$ is a theorem of LM_P. Consider a world w in model M such that $M, w \vDash \mathcal{P}(\phi/\psi)$. Def. 4.6 gives us $(\|\phi\|, \|\psi\|) \in N_{\mathcal{P}}(w)$, and (i) from Def. 4.5 gives us that $(\|\phi\|, W) \in N_{\mathcal{F}}(w)$ or $(\|\neg\phi\|, W) \in N_{\mathcal{O}}(w)$. Since $W = \|\top\|$, we have that $M, w \vDash \mathcal{F}(\phi/\top)$ or $M, w \vDash \mathcal{O}(\neg\phi/\top)$. Therefore, $M, w \vDash \mathcal{P}(\phi/\psi) \to (\mathcal{F}(\phi/\top) \vee \mathcal{O}(\neg\phi/\top))$. □

Theorem 4.9 (Completeness) *If a formula ϕ is valid, then ϕ is a theorem of LM_P.*

Proof. We use the method of canonical models from [13]. First, we define the canonical model M^c, in such a way that for each formula ϕ and world w, $M^c, w \vDash \phi$ iff $\phi \in w$. The formulas true in all worlds of M^c are, then, exactly the theorems of LM_P. M^c is not necessarily an LM_P-model. The universal modality $\boxed{\mathsf{U}}$ is axiomatized by S5, which is canonical for the equivalence relation, i.e. $R^c_{\boxed{\mathsf{U}}} \subseteq W \times W$ (see [10]). For the global modality, the required property is $R^c_{\boxed{\mathsf{U}}} = W \times W$. Thus, as done in [8], we introduce a submodel M^* of the canonical model M^c, and show that M^* is an LM_P-model. M^* is then used to establish completeness.

The canonical model $M^c = \langle W^c, R^c_{\boxed{\mathsf{U}}}, N^c_{\mathcal{O}}, N^c_{\mathcal{P}}, N^c_{\mathcal{F}}, V^c \rangle$ for LM_P is defined as follows. W^c is the set of all LM_P-maximally consistent sets of formulas. Let $(Y, Z) \in N^c_{\mathcal{O}}(w)$ iff $Y \neq W^c$ and there is a formula $\mathcal{O}(\phi/\psi) \in w$ such that $\{w_j \in W^c \mid \phi \in w_j\} \subseteq Y$ and $\{w_j \in W^c \mid \psi \in w_j\} = Z$. Then, let $(Y, Z) \in N^c_{\mathcal{P}}(w)$ iff there is a formula $\mathcal{P}(\phi/\psi) \in w$ such that $Y = \{w_j \in W^c \mid \phi \in w_j\}$

and $\{w_j \in W^c \mid \psi \in w_j\} = Z$. Furthermore, let $(Y, Z) \in N^c_{\mathcal{F}}(w)$ iff $Y \neq \emptyset$ and there is a formula $\mathcal{F}(\phi/\psi) \in w$ such that $Y \subseteq \{w_j \in W^c \mid \phi \in w_j\}$ and $\{w_j \in W^c \mid \psi \in w_j\} = Z$. Lastly, $w \in V^c(p)$ iff $p \in w$. We will use the following shorthand throughout the proof $\|\phi\|^c = \{w \in W^c \mid \phi \in w\}$.

To show that our canonical model satisfies the restrictions of Def. 4.5, we outline the case of (vi). The same strategy can be adopted for the other cases.

(vi) If $(X, Z) \in N^c_{\mathcal{P}}(w)$, then $(\overline{X}, Z) \notin N^c_{\mathcal{O}}(w)$. To see why, consider $(X, Z) \in N^c_{\mathcal{P}}$ for some $w \in W^c$ and $X, Z \subseteq W^c$. Note that there is a formula $\mathcal{P}(\phi/\psi) \in w$ where $\|\phi\|^c = X$ and $\|\psi\|^c = Z$. By axiom P2c, we have $\mathcal{O}(\neg\phi/\psi) \notin w$. It might be the case that $(\overline{X}, Z) \in N^c_{\mathcal{O}}(w)$ if there is a χ such that $\|\chi\|^c \subseteq \|\neg\phi\|^c$ and $\mathcal{O}(\chi/\psi) \in w$. However, since $\boxed{\scriptstyle U}(\chi \to \neg\phi)$, by Ax1, we have $\mathcal{O}(\neg\phi/\psi) \in w$ contradicting P2c. Thus, $(\overline{X}, Z) \notin N^c_{\mathcal{O}}(w)$.

We show, by induction on ϕ, that $M^c, w \vDash \phi$ iff $\phi \in w$. The base case is clear: $M^c, w \vDash p$ implies $p \in w$ by definition. For the inductive case, we consider only $\mathcal{P}(\phi/\psi)$, as the classical connectives are straightforward, and $\mathcal{O}(\phi/\psi)$ and $\mathcal{F}(\phi/\psi)$ are done similarly. If $M^c, w \vDash \mathcal{P}(\phi/\psi)$, then $(\|\phi\|^c, \|\psi\|^c) \in N^c_{\mathcal{P}}(w)$. By the canonical model, there is a formula $\mathcal{P}(\theta_1/\theta_2) \in w$ such that $\|\phi\|^c = \|\theta_1\|^c$ and $\|\psi\|^c = \|\theta_2\|^c$. By axioms P4ab, we have that $\mathcal{P}(\phi/\psi) \in w$.

Our model M^c satisfies $R^c_{\boxed{\scriptstyle U}} \subseteq W^c \times W^c$. However, since $\boxed{\scriptstyle U}$ represents the global modality, it is necessary that $R^c_{\boxed{\scriptstyle U}} = W^c \times W^c$. To meet this requirement, we generate a submodel M^* of M^c, and show that its relation $R^*_{\boxed{\scriptstyle U}}$ satisfies $R^*_{\boxed{\scriptstyle U}} = W^* \times W^*$ for some $W^* \subseteq W^c$. We then prove that M^* is an LM_P-model, and utilize this model to demonstrate completeness. To construct $M^* = \langle W^*, R^*_{\boxed{\scriptstyle U}}, N^*_{\mathcal{O}}, N^*_{\mathcal{F}}, N^*_{\mathcal{P}}, V^* \rangle$ we begin by selecting a world $w \in W^c$. Its domain W^* is defined as follows: $W^* = \{v \in W^c \mid \text{for all } \boxed{\scriptstyle U}\phi \in w,\ \phi \in v\}$. The relation $R^*_{\boxed{\scriptstyle U}}$ as $R^*_{\boxed{\scriptstyle U}} = R^c_{\boxed{\scriptstyle U}} \cap (W^* \times W^*)$. As described in [10], it can be easily shown that $R^*_{\boxed{\scriptstyle U}} = W^* \times W^*$, which is our required property. Then, $V^*(p) = V^c(p) \cap W^*$, and the neighborhood functions are defined as follows: $N^*_{\chi}(w) = \{(X, Y) \mid (X', Y') \in N^c_{\chi}(w), X = X' \cap W^*, Y = Y' \cap W^*\}$ for $\chi \in \{\mathcal{O}, \mathcal{P}, \mathcal{F}\}$. By a simple induction on the complexity of ϕ, it follows that $\|\phi\|^* = \{w \in W^* \mid \phi \in w\} = \|\phi\|^c \cap W^*$. We can show that each neighbourhood restriction is satisfied by M^*, and that M^* is thus a LM_P-model. We show the case for (i), the other cases being similar.

(i) If $(X, Y) \in N^*_{\mathcal{P}}(w)$ then $(X, W^*) \in N^*_{\mathcal{F}}(w)$ or $(\overline{X}, W^*) \in N^*_{\mathcal{O}}(w)$. To see why, consider $(X, Y) \in N^*_{\mathcal{P}}(w)$ for some $w \in W^*$. Then, by definition of the submodel M^*, it follows that $X = X' \cap W^*$ and $Y = Y' \cap W^*$ for some $(X', Y') \in N^c_{\mathcal{P}}(w)$. Since M^c is an LM_P model, we know that $(X', W^c) \in N^c_{\mathcal{F}}(w)$ or $(\overline{X'}, W^c) \in N^c_{\mathcal{O}}(w)$. Thus, $(X' \cap W^*, W^c \cap W^*) = (X, W^*) \in N^*_{\mathcal{F}}(w)$ or $(\overline{X'} \cap W^*, W^c \cap W^*) = (\overline{X}, W^*) \in N^*_{\mathcal{O}}(w)$.

Lastly, we have that for each $w \in W^*$, $M^*, w \vDash \phi \leftrightarrow M^c, w \vDash \phi$ by induction on the complexity of ϕ, and therefore $M^*, w \vDash \phi \leftrightarrow \phi \in w$. □

Lemma 4.10 (Consistency) *The logic LM_P is consistent.*

Proof. We exhibit a LM_P-model M in which all LM_P axioms hold but there is one formula that does not. Let $M = \langle W, N_{\mathcal{O}}, N_{\mathcal{F}}, N_{\mathcal{P}}, V \rangle$, where $W =$

$\{w_1, w_2\}$, $N_{\mathcal{O}}(w_i) = \{(\{w_1\}, \{w_2\})\}$, $N_{\mathcal{P}}(w_i) = N_{\mathcal{F}}(w_i) = \emptyset$ for $i \in \{1, 2\}$, and $V(p) = \{w_1\}$, $V(q) = \{w_2\}$.

We show that axiom P2b holds. We have $M, w_i \vDash \mathcal{O}(p/q)$, since $(\|p\|, \|q\|) \in N_{\mathcal{O}}(w_i)$, and $M, w_i \nvDash \mathcal{P}(p/q)$. The model similarly satisfies axiom P2a, P2c, Ax3 and Ax4. It trivially satisfies all remaining axioms, since they are implications and the antecedent is false. Since $M, w_i \nvDash \mathcal{P}(p/q)$, there is a formula that does not hold in the model and therefore LM_P is consistent. □

5 Deontic Paradoxes in Mīmāṃsā

To analyze the behaviour of LM_P we use as benchmarks the main deontic paradoxes[6] involving permission: the free choice inference [42], Ross' paradox [37] and the paradox of the privacy act [22]. As demonstrated below, LM_P behaves well with respect to them.

5.1 The Free Choice Inference

It is plausible to say that "you may have coffee or tea" implies that you may have a coffee and you may have a tea (though possibly not both at once). This very intuitive principle, first mentioned in [42], is known as the free choice inference (FCI) and is formalized in SDL as $\mathcal{P}(\phi \vee \psi) \to \mathcal{P}(\phi)$. The paradoxical consequences of accepting FCI have been widely discussed in deontic logic, see, e.g. [21,5,12,16]. Among them, as demonstrated in [21], SDL with (FCI) derives (i) $\mathcal{O}(\phi) \to \mathcal{O}(\phi \wedge \psi)$, (ii) $\mathcal{O}(\phi) \to \mathcal{P}(\psi)$, (iii) $\mathcal{P}(\phi) \to \mathcal{P}(\psi)$ and (iv) $\mathcal{P}(\phi) \to \mathcal{P}(\phi \wedge \psi)$. As a special instance of (iii), we get (v) $\mathcal{P}(\phi) \to \mathcal{P}(\bot)$, which is a particularly undesirable consequence in Mīmāṃsā, where permitted actions should be possible, as shown by Lemma 4.4. As a result, we modify the free choice inference in LM_P to ensure that every inferred permission corresponds to a feasible action:

$$\mathcal{P}(\phi \vee \psi/\theta) \wedge \Diamond\mkern-12.5mu{\scriptscriptstyle\mathrm{U}}\mkern6mu \phi \to \mathcal{P}(\phi/\theta). \qquad (\mathrm{FCI}\Diamond\mkern-12.5mu{\scriptscriptstyle\mathrm{U}}\mkern6mu)$$

We demonstrate that the (dyadic variant of) (i)-(v) cannot be derived in LM_P in presence of FCI$\Diamond\mkern-12.5mu{\scriptscriptstyle\mathrm{U}}\mkern6mu$. We start by establishing a sufficient condition for FCI$\Diamond\mkern-12.5mu{\scriptscriptstyle\mathrm{U}}\mkern6mu$ to hold in an LM_P-model.

Lemma 5.1 *Let $M = \langle W, N_{\mathcal{O}}, N_{\mathcal{P}}, N_{\mathcal{F}}, V\rangle$ be an LM_P-model, and consider non-empty $X, Y, Z \subseteq W$. For all $w \in W$, if $X \subseteq Y$ and $(Y, Z) \in N_{\mathcal{P}}(w)$ implies $(X, Z) \in N_{\mathcal{P}}(w)$, then $M, w \vDash$FCI$\Diamond\mkern-12.5mu{\scriptscriptstyle\mathrm{U}}\mkern6mu$.*

Proof. Assume $M, w \vDash \mathcal{P}(\phi \vee \psi/\theta) \wedge \Diamond\mkern-12.5mu{\scriptscriptstyle\mathrm{U}}\mkern6mu \phi$. Then, $(\|\phi\| \cup \|\psi\|, \|\theta\|) \in N_{\mathcal{P}}(w)$. Since $\|\phi\| \subseteq \|\phi\| \cup \|\psi\|$ and $\|\phi\| \neq \emptyset$ (by $M, w \vDash \Diamond\mkern-12.5mu{\scriptscriptstyle\mathrm{U}}\mkern6mu \phi$), we have that $(\|\phi\|, \|\theta\|) \in N_{\mathcal{P}}(w)$ and thus $M, w \vDash \mathcal{P}(\phi/\theta)$. □

The example below exhibits an LM_P-model that satisfies FCI$\Diamond\mkern-12.5mu{\scriptscriptstyle\mathrm{U}}\mkern6mu$ but not the unwanted consequences (i)-(v).

[6] Although called paradoxes, they are intended here in a broad sense as (un)derivable theorems that are counter-intuitive in a common-sense reading.

Example 5.2 Let $M = \langle W, N_{\mathcal{O}}, N_{\mathcal{P}}, N_{\mathcal{F}}, V\rangle$ be the LM_P-model such that $W = \{w_1, w_2, w_3\}$, $V(p) = \{w_1\}$, $V(q) = \{w_2\}$, $V(r) = \{w_3\}$, $N_{\mathcal{P}}(w_i) = \{(V(q), V(r))\}$, $N_{\mathcal{O}}(w_i) = \{(X, Y) \mid V(p) \subseteq X, X \neq W, Y = V(r)\}$, $N_{\mathcal{F}}(w_i) = \{(V(p), W)\}$ for $i \in \{1, 2, 3\}$. FCI$\Diamond^{u}$ is true in all $w_i \in W$, by Lem. 5.1. We show that M does not satisfy (i)-(v).

For (i), we see that $M, w_i \vDash \mathcal{O}(p/r)$ and $M, w_i \nvDash \mathcal{O}(p \wedge q/r)$. For (ii), $M, w_i \vDash \mathcal{O}(p/r)$ and $M, w_i \nvDash \mathcal{P}(r/r)$. For (iii), we have $M, w_i \vDash \mathcal{P}(q/r)$ and $M, w_i \nvDash \mathcal{P}(p/r)$. For (iv), we have that $M, w_i \vDash \mathcal{P}(q/r)$ and $M, w_i \nvDash \mathcal{P}(p \wedge q/r)$. Lastly, for (v), we have $M, w_i \vDash \mathcal{P}(q/r)$ and $M, w_i \nvDash \mathcal{P}(\bot/r)$.

Remark 5.3 The undesirable consequences (i)-(v) can be derived in SDL using instances of obligation implies permission (aka axiom D), interdefinability between the deontic operators, and monotonicity of permission. These principles do not hold in LM_P. Nonetheless LM_P cannot get rid of *all possible* unwanted results. To elaborate: while the undesirable inferences regarding obligation (i.e., (i) and (ii)), and impossible actions (i.e., (v)) are blocked even when an unrelated action ψ is possible, in the presence of FCI$\Diamond^{u}$, due to axiom P4b the statement $\mathcal{P}(\phi/\theta) \wedge \Diamond^{u}(\phi \wedge \psi) \rightarrow \mathcal{P}(\phi \wedge \psi/\theta)$ can be derived in LM_P. This debatable statement asserts that if ϕ is permitted, so is $\phi \wedge \psi$, for any compatible action ψ.

5.2 Ross' paradox

Ross' paradox [37] is a frequently debated issue. Introduced as a paradox for obligation, it states that the obligation to mail a letter implies the obligation to mail the letter or burn it. Here we consider its version for permission ("the permission to mail the letter implies the permission to mail or burn the letter"), formalized as the following valid formula in SDL

$$\mathcal{P}(\phi) \rightarrow \mathcal{P}(\phi \vee \psi).$$

The prima facie version of this paradox does not apply to permissions in Mīmāṃsā, because all commands in Mīmāṃsā have only one action as their argument. Moreover, the consequences of the paradox can be avoided even if we consider the all-things-considered deontic situation. In fact, as discussed in Section 3, unconditional permissions do not exist in Mīmāṃsā and thus the dyadic version of the paradox is the following:

$$\mathcal{P}(\phi/\theta) \rightarrow \mathcal{P}(\phi \vee \psi/\theta).$$

This formula is not derivable in LM_P, as shown by the following countermodel:

Let $M = \langle W, N_{\mathcal{O}}, N_{\mathcal{P}}, N_{\mathcal{F}}, V\rangle$ be a LM_P-model, such that $W = \{w_1, w_2, w_3\}$, $V(p) = \{w_1\}$, $V(q) = \{w_2\}$, $V(r) = \{w_3\}$, $N_{\mathcal{P}}(w_i) = \{(V(p), V(r))\}$, $N_{\mathcal{F}}(w_i) = \{(V(p), W)\}$ and $N_{\mathcal{O}}(w_i) = \emptyset$ for $i \in \{1, 2, 3\}$. Note that the neighborhood function of prohibition is not empty in order to satisfy condition (i) stated in Def. 4.5. We see that $(V(p), V(r)) \in N_{\mathcal{P}}(w_i)$, but $(V(p) \cup V(q), V(r)) \notin N_{\mathcal{P}}(w_i)$. Thus $M, w_i \vDash \mathcal{P}(p/r)$ while $M, w_i \nvDash \mathcal{P}(p \vee q/r)$.

Remark 5.4 Ross' paradox does not appear in LM_P as Mīmāṃsā permission is not monotonic in the first argument. If we were to derive $\mathcal{P}(\text{mail} \lor \text{burn}/\theta)$ from $\mathcal{P}(\text{mail}/\theta)$ for some θ, then we would need to have a pre-existing command $\mathcal{F}(\text{burn}/\top)$ or $\mathcal{O}(\neg\text{burn}/\top)$ (cf. Remark 4.3). This is impossible if such a pre-existing prohibition or negative obligation is not available.

5.3 The Paradox of the Privacy Act

Introduced in [22], this paradox consists of a privacy act containing the norms:

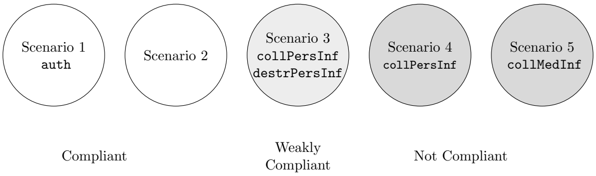
(i) The collection of personal information is forbidden unless acting on a court order authorising it.
(ii) The destruction of illegally collected personal information before accessing it is a defence against the illegal collection of personal data.
(iii) The collection of medical information is forbidden unless the entity collecting the medical information is permitted to collect personal information.

To properly assess this act, we need to consider five distinct scenarios as all other possible scenarios are variations of these. We refer to these as Scenarios 1-5. Scenario 1 involves a court order that authorizes the collection of personal data. Regardless of whether the data is ultimately collected or not, this scenario is compliant with the privacy act. Scenario 2, where a court has not authorized the collection of data and neither personal nor medical data is collected, is compliant as well. Scenario 3, where personal data is collected illegally but is compensated by its destruction, is called 'weakly compliant'. Lastly, there are two non-compliant situations: Scenario 4, involving the unauthorized collection of personal data, and Scenario 5, involving the unauthorized collection of medical data.

While SDL can formalize the norms (i)-(iii) in a consistent way, it derives a contradiction when considering the compliant Scenarios 1 and 2. For, by formalizing (i) as $\mathcal{F}(\mathit{collPersInf})$ and $\mathit{auth} \to \mathcal{P}(\mathit{collPersInf})$, when auth is true (as in Scenario 1), we derive $\mathcal{P}(\mathit{collPersInf})$, contradicting $\mathcal{F}(\mathit{collPersInf})$.

This contradiction is prevented in LM_P. We formalize the norms (i)-(iii) in the following way: (i) is $\mathcal{F}(\mathit{collPersInf}/\top)$ and $\mathcal{P}(\mathit{collPersInf}/\mathit{auth})$. Norm (ii) represents a contrary-to-duty obligation (see e.g. [35]) since the violation of collecting personal data must be compensated by its destruction, and is formalized as $\mathcal{O}(\mathit{destrPersInf}/\mathit{collPersInf})$. Lastly, (iii) is formalized as $\mathcal{F}(\mathit{collMedInf}/\top)$ and $\mathcal{P}(\mathit{collPersInf}/X) \to \mathcal{P}(\mathit{collMedInf}/X)$ for any X, since the permission of collecting medical data depends on the condition X of the permission for collecting personal data.

We show that LM_P is suitable to model the privacy act, by giving a model where all norms (i)-(iii) holds, and each world represents one of the scenarios without contradictions: $M = \langle W, N_{\mathcal{O}}, N_{\mathcal{P}}, N_{\mathcal{F}}, V\rangle$, where: $W = \{w_i \mid 1 \leq i \leq 5\}$, $\|\mathit{collPersInf}\| = \{w_3, w_4\}$, $\|\mathit{destrPersInf}\| = \{w_3\}$, $\|\mathit{auth}\| = \{w_1\}$, $\|\mathit{collMedInf}\| = \{w_5\}$, $N_{\mathcal{F}}(w_i) = \{(X, Y) \mid X \neq \emptyset, X \subseteq \|\mathit{collPersInf}\|, Y = W\} \cup \{(U, Z) \mid U \neq \emptyset, U \subseteq \|\mathit{collMedInf}\|, Z = W\}$, $N_{\mathcal{O}}(w_i) = \{(X, Y) \mid \|\mathit{destrPersInf}\| \subseteq X, Y = \|\mathit{collPersInf}\|\}$, and $N_{\mathcal{P}}(w_i) = \{(\|\mathit{collPersInf}\|, \|\mathit{auth}\|), (\|\mathit{collMedInf}\|, \|\mathit{auth}\|)\}$. The picture below illustrates the model.



Note that the exception-based definition of permission in LM_P is well-suited for the formalization of the privacy act, which considers permissions as exceptions to prohibitions.

Remark 5.5 The paradox is resolved in LM_P by the use of dyadic deontic operators. In contrast to SDL's monadic operators, LM_P indeed enables the derivation of context-dependent prohibitions, permissions, and obligations, accommodating changing situations, and thus allowing, e.g., the formulas $\mathcal{F}(collPersInf/\top)$ and $\mathcal{P}(collPersInf/auth)$ to be true simultaneously.

6 Conclusions

Mīmāṃsā provides a treasure trove of more than 2,000 years worth of deontic investigations, including the application of deontic principles in juridical contexts and problems. In this article, we have analyzed the notion of permission in Mīmāṃsā, and formalized its properties by transforming relevant *nyāya*s (identified, translated from Sanskrit and interpreted) into suitable Hilbert axioms. The resulting deontic operator has been added to the logic of Mīmāṃsā as discussed in [8], and a sound and complete semantics has been provided. We have analyzed the behavior of the new permission operator using an established method in the deontic logic literature, which involves confronting it with deontic paradoxes, and found out that the resulting operator behaves well w.r.t. the considered paradoxes.

One might wonder whether the command we are discussing can be meaningfully described as permission at all. In fact, the term 'permission' in Euro-American philosophy or in Deontic Logic is strongly polysemic, covering, among others, acts that are not normed as well as acts that were previously prohibited and are now permitted, and even rights. Philosophers of the Mīmāṃsā school, by contrast, adopt the standard Sanskrit terms for permission (*anujñā* and *anumati*), but focus on only one aspect among the ones mentioned above, and use different terms for the others (for instance, *adhikāra* comes close to rights, see [18]). Using the term 'permission' thus highlights a single shared aspect and suggests a way out of the polysemy of 'permissions'.

Overall, this paper introduces and formalizes the concept of permission in Mīmāṃsā, contributes to the ongoing development of deontic logic, and sheds light on the importance of considering permission in normative reasoning.

There is still a missing component to capture the essence of Mīmāṃsā permission. As discussed in Section 3, while a certain condition may render a generally prohibited action permissible under specific circumstances, Mīmāṃsā still encourages avoiding such action whenever possible. To address this, we aim to incorporate in LM_P the *Ceteris Paribus* preference (as e.g. in [7,31]) as future work. Specifically, we plan to compare two scenarios with identical obligations and prohibitions, but where the preference of the world depends on the fulfilled permissions.

Acknowledgement

Work partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 101034440.

References

[1] Alchourrón, C. E. and E. Bulygin, *Permission and permissive norms*, Theorie der Normen. Festgabe für Ota Weinberger zum 65. Geburtstag (1984), pp. 349–371.

[2] Alchourrón, C. E., *Permissory Norms and Normative Systems (1984/86/2012)*, in: *Essays in Legal Philosophy*, Oxford University Press, 2015 .

[3] Anglberger, A. J. J., H. Dong and O. Roy, *Open reading without free choice*, in: F. Cariani, D. Grossi, J. Meheus and X. Parent, editors, *Deontic Logic and Normative Systems* (2014), pp. 19–32.

[4] Åqvist, L., *Deontic logic*, in: D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic: Volume II*, Springer, Dordrecht, 1984 pp. 605–714.

[5] Asher, N. and D. Bonevac, *Free choice permission is strong permission*, Synthese **145** (2005), pp. 303–323.

[6] Barker, C., *Free choice permission as resource-sensitive reasoning*, Semantics and Pragmatics **3** (2010), pp. 1–38.

[7] Benthem, J. v., P. Girard and O. Roy, *Everything else being equal: A modal logic for ceteris paribus preferences*, Journal of philosophical logic **38** (2009), pp. 83–125.

[8] Berkel, K. v., A. Ciabattoni, E. Freschi, F. Gulisano and M. Olszewski, *Deontic paradoxes in mīmāṃsā logics: There and back again*, Journal of Logic, Language and Information (2022), pp. 1–44.

[9] Berkel, K. v. and T. Lyon, *The varieties of ought-implies-can and deontic stit logic*, in: F. Liu, A. Marra, P. Portner and F. V. D. Putte, editors, *Proceedings of DEON 2021*, 2021 .

[10] Blackburn, P., M. d. Rijke and Y. Venema, "Modal Logic," Cambridge Tracts in Theoretical Computer Science, Cambridge University Press, 2001.

[11] Bouvier, J., "A Law Dictionary," Childs and Peterson, Philadelphia, 1856.

[12] Broersen, J. and L. van der Torre, *Ten problems of deontic logic and normative reasoning in computer science*, Lectures on Logic and Computation: ESSLLI 2010, Selected Lecture Notes (2012), pp. 55–88.

[13] Chellas, B. F., "Modal Logic," Cambridge University Press, Cambridge, 1980.

[14] Ciabattoni, A., E. Freschi, F. A. Genco and B. Lellmann, *Mīmāṃsā deontic logic: Proof theory and applications*, in: H. De Nivelle, editor, *Automated Reasoning with Analytic Tableaux and Related Methods*, Springer International Publishing, Cham, 2015 pp. 323–338.

[15] Danielsson, S., "Preference and Obligation," Filosofiska Färeningen, Uppsala, 1968.

[16] Dignum, F., J.-J. C. Meyer and R. Wieringa, *Contextual permission. a solution to the free choice paradox*, in: *Second International Workshop on Deontic Logic in Computer Science*, 1994, pp. 107–135.

[17] Echave, D. T., M. E. Urquijo and R. Guibourg, "Lógica, proposición y norma," Astrea, Buenos Aires, 1980.

[18] Freschi, E., *Mīmāṃsā and dharmaśāstra sources on permissions*, in: A. Cerulli and P. A. Maas, editors, *Festschrift for Dominik Wujastyk*, forthcoming in 2024 .

[19] Freschi, E., A. Ciabattoni, F. A. Genco and B. Lellmann, *Understanding prescriptive texts: Rules and logic as elaborated by the Mīmāṃsā school*, Journal of World Philosophies **2** (2017), pp. 47–66.

[20] Freschi, E. and M. Pascucci, *Deontic concepts and their clash in mīmāṃsā: Towards an interpretation*, Theoria **87** (2021), pp. 659–703.

[21] Gabbay, D., L. Gammaitoni and X. Sun, *The paradoxes of permission an action based solution*, Journal of Applied Logic **12** (2014), pp. 179–191.

[22] Governatori, G., *Thou shalt is not you will*, in: *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15 (2015), p. 63–68.

[23] Gustafsson, J. E., *Permissibility is the only feasible deontic primitive*, Philosophical Perspectives **34** (2020), pp. 117–133.

[24] Hansson, B., *An analysis of some deontic logics*, Noûs **3** (1969), pp. 373–398, reprinted in [28, pp. 121-147].

[25] Hansson, S. O., *Preference-based deontic logic (pdl)*, Journal of Philosophical Logic (1990), pp. 93–122.

[26] Hansson, S. O., *The varieties of permission*, in: D. M. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of deontic logic and normative systems*, College Publications, London, 2013 pp. 195–240.

[27] Hansson, S. O., *In defence of deontic diversity*, Journal of Logic and Computation **29** (2015), pp. 349–367.
URL `https://doi.org/10.1093/logcom/exv057`

[28] Hilpinen, R., editor, "Deontic Logic," Reidel, Dordrecht, 1971.

[29] Lellmann, B., F. Gulisano and A. Ciabattoni, *Mīmāṃsā deontic reasoning using specificity: a proof theoretic approach*, Artificial Intelligence and Law **29** (2021), pp. 351–394.

[30] Lewis, D., "Counterfactuals," Blackwell, Oxford, 1973.

[31] Loreggia, A., E. Lorini and G. Sartor, *Modelling ceteris paribus preferences with deontic logic*, Journal of Logic and Computation **32** (2022), pp. 347–368.

[32] Makinson, D. and L. van der Torre, *Permission from an input/output perspective*, Journal of philosophical logic **32** (2003), pp. 391–416.

[33] McNamara, P., *Deontic logic*, in: D. M. Gabbay and J. Woods, editors, *Logic and the Modalities in the Twentieth Century*, Handbook of the History of Logic **7**, North-Holland, 2006 pp. 197–288.

[34] Olszewski, M., X. Parent and L. van der Torre, *Input/output logic with a consistency check-the case of permission.*, in: F. Liu, A. Marra, P. Portner and F. V. D. Putte, editors, *Proceedings of DEON 2021*, 2021, pp. 358–375.

[35] Prakken, H. and M. Sergot, *Contrary-to-duty obligations*, Studia Logica **57** (1996), pp. 91–115.

[36] Rescher, N. and A. R. Anderson, *Conditional permission in deontic logic*, Philosophical Studies **13** (1962), pp. 1–8.

[37] Ross, A., *Imperatives and logic*, Philosophy of Science **11** (1944), p. 30–46.

[38] Stern, R., *Does 'ought' imply 'can'? and did kant think it does?*, Utilitas **16** (2004), pp. 42–61.

[39] van Fraassen, B., *The logic of conditional obligation*, J. of Phil. Logic **1** (1972), pp. 417–438.

[40] von Wright, G. H., *Deontic logic*, Mind **60** (1951), pp. 1–15.

[41] von Wright, G. H., "Norm and Action," Routledge and Kegan Paul, London, 1963.

[42] von Wright, G. H., "An Essay in Deontic Logic and the General Theory of Action: With a Bibliography of Deontic and Imperative Logic," Amsterdam: North-Holland Pub. Co., 1968.

[43] von Wright, G. H., *Deontic logic and the theory of conditions*, in: *Deontic Logic: Introductory and Systematic Readings*, Springer Netherlands, 1970 pp. 159–177.

[44] Zylberman, A., *Moral rights without balancing*, Philosophical Studies **179** (2022), p. 549–569.

Beyond reasons and obligations: A dual-role approach to reasons and supererogation

Aleks Knoks [1]

University of Luxembourg
2, avenue de l'Université
L-4365 Esch-sur-Alzette

David Streit [1]

University of Luxembourg
2, avenue de l'Université
L-4365 Esch-sur-Alzette

Abstract

Dual-role approaches to reasons say, roughly, that reasons can relate to actions in two fundamentally different ways: they can either require conformity, or justify an action without requiring that it be taken. This paper develops a formal dual-role approach, combining ideas from defeasible logic and practical philosophy. It then uses the approach to shed light on the phenomenon of supererogation and resolve a well-known puzzle about supererogation, namely, Horton's All or Nothing Problem.

Keywords: reasons, dual-role approaches, defeasible logic, supererogation, all or nothing problem

1 Introduction

This paper has two goals. The first is to capture the core idea behind what we call *dual-role approaches* to reasons in a simple defeasible logic—in the style of the logic presented in [18]. The idea in question is, roughly, that the normative forces associated with reasons are of two fundamentally different kinds: some reasons for action *require* conformity, while others *justify* actions without requiring that they be taken. The second goal is to apply the resulting formal dual-role approach to the phenomenon of supererogation and to develop a unified response to the puzzles surrounding it.

[1] The first author acknowledges the support of the Fonds National de la Recherche Luxembourg (FNR) through the projects DELIGHT (OPEN O20/14776480). Both authors acknowledge the travel support from the same funding institution through the project INTEGRAUTO (INTER/AUDACE/21/16695098). We also thank our three anonymous reviewers for their insightful comments. The second reviewer's generous set of comments was particularly helpful.

The remainder of this paper is structured as follows. Section 2 provides a quick survey of the relevant literature on reasons, dual-role approaches, and formal work. Section 3 sets up the formal model. Section 4 shifts the focus to supererogation and explains how the dual-role approach that comes with the model responds to the basic challenge supererogation poses. Section 5 discusses a further puzzle of supererogation, namely, the *All or Nothing Problem*, [14]. Section 6 discusses related (formal) work on supererogation. The concluding Section 7 is followed by an appendix that contains the proofs of the most important results.

2 Reasons and dual-role approaches

Practical normative reasons are standardly characterized as considerations that count in favor or against actions. [2] Schroeder [41] helpfully points out the three "marks" that are characteristic of such reasons: they compete against each other, they are act-oriented, and they are the sorts of considerations that one can act for. The notion has become a mainstay of practical philosophy, where it is routinely made use of in answering various normative and metanormative questions. This is taken to the extreme in the *reasons-first program* which holds, roughly, that the notion of reasons is basic and that all other normative notions are to be analyzed in terms of it. [3]

For our purposes, two more recent developments in the literature on reasons will be particularly important. The first is the formal work on reasons, and, in particular, Horty's default logic-based model of the way reasons interact to support oughts. [4] Since the publication of *Reasons as Defaults* [18], this model has been extended in several ways and applied to many new problems, even finding a path into a more orthodox (that is, nonformal) monograph on reasons. [5] Other frameworks have been used to model reasons too—see, e.g., [6] and [7]—but default logic and defeasible logics more generally have been more influential. The second body of literature crucial for our interests develops what we call *dual-role approaches* to reasons. [6] Lately, several authors—most notably, Gert [8], [9], [10] and Greenspan [11], [12]—have argued that we need to distinguish two fundamentally different dimensions in the normative forces associated with reasons. [7] Thus, Gert discusses "requiring and justifying strengths" of reasons: the requiring strength is said to ground potential criticism and, through that,

[2] See, e.g., [34], [37], [38], [40], [44]. Whenever we say *reasons*, we always mean *normative* reasons, as opposed to explanatory or motivational reasons—see [1], for a discussion.

[3] The locus classicus here is Scanlon [38]. But see also, e.g., [34], [37], [41].

[4] We use the terms *ought* and *oughts* to refer to conclusions about what we ought to do.

[5] See, e.g., [15], [28], [32], and [41, Chs. 4.4–5].

[6] We borrow the term from [29].

[7] Both Gert and Greenspan urge to draw the distinction since it allows one to resolve various foundational issues in practical philosophy. For instance, Gert [8] shows how it can be used to the benefit of certain moral theories which, without the distinction, allow for cases in which the agent is forced to choose between an irrational moral action and an immoral rational one.

require conformity, while the justifying strength is said to ground answers to potential criticism and, thereby, *justify* nonconformity—see, e.g., [9, p. 541]. Importantly, some reasons are meant to be "purely justifying", meaning that they possess only the latter type of strength. In a similar vein, Greenspan [11,12] discusses "negative reasons" which count against an action and, without sufficient counterbalancing reasons, ".. [ground] a requirement to take some alternative option.." [11, p. 387] and "positive reasons" which count in favor of acts. Greenspan takes purely positive reasons to "ground at most only a recommendation". They "do not compel, but instead are optional, rendering an option eligible for choice, or justifying it, without requiring it" [11, p. 389]. [8] Gert's and Greenspan's views differ in details, but these won't matter for our purposes. Our main takeaway is their (common) core insight: reasons can relate to actions in two fundamentally different ways: They can have requiring force or (merely) justify an action. [9] The goal of the next section is to make this precise by expressing the core insight of dual-role approaches in a defeasible logic and combining the two strands found in the literature. The only other published attempt at formalizing dual-role approaches that we are aware of is due to Mullins [29]. While Mullins builds on Horty's model, like we do, his formalization differs from ours in several important respects. We compare our approach to his in Section 6.

3 The formal model

Let's start with a simple scenario:

> **Save One or Two**. Alice and Bob are trapped in a collapsing building. You can easily and without costs to yourself save one of them. You can also save both, but that would involve serious harm to you: you would lose your legs. [10]

Notice the three reasons that are particularly salient in this scenario: the fact that Alice will die, unless you save her; the fact that Bob will die, unless you save him; and the fact that you will lose your legs if you save both. Notice too, that all of the following judgments seem very intuitive: you *have to* save either Alice or Bob (we'd blame you if you walked away); it's not the case that you *have to* save both (we wouldn't blame you for deciding to keep your legs); but

[8] It pays noting that, in the literature on reasons, the terms *positive* and *negative reasons* are often applied to, respectively, reasons that count in favor of an action and those that count against—see, e.g., [37]. Clearly, this is very different from the way Greenspan uses these terms. To avoid confusion, we adopt Gert's terminology.

[9] Many other authors have drawn similar distinctions. This includes Dancy's [5] distinction between "enticing" and "peremptory" reasons, Parfit's [34] distinction between "partial" and "impartial" reasons, Portmore's [35] distinction between "moral" and "nonmoral" reasons, and Muñoz's [30] distinction between reasons and "prerogatives". The idea is always that we can distinguish two different dimensions in the way reasons—or reasons and considerations that aren't reasons—relate to actions.

[10] The scenario comes from [31].

were you to save both, your action would be highly admirable.[11]

We will now devise a formal notation that is just rich enough for a dual-role analysis of this scenario. As background, we assume the language of propositional logic with the standard connectives (including $\bot$), and we let the customary symbol $\vdash$ stand for classical logical consequence. Thus, we can use the propositional letters A, B, and L to express the propositions, respectively, that you save Alice, that you save Bob, and that you lose your legs. The constraint that you can't save Alice and Bob, as well as keep your legs, can be expressed as the material conditional $(A\&B) \supset L$. Extending the language slightly, we allow for formulas of the form $!X$ and read them as saying that there is a reason supporting the proposition expressed by X. What $!A$, $!B$, and $!\neg L$, then, say is, respectively, that there's a reason supporting your saving Alice, that there's a reason supporting your saving Bob, and that there's a reason supporting your not losing your legs.[12] For our purposes, it is not important to explicitly represent the reasons that ground such formulas as $!A$, $!B$, and $!\neg L$. In all the cases we will discuss, it won't matter what these reasons are exactly. What's more, we won't encounter any cases where the fact that a proposition is supported by multiple different reasons can make a difference for its analysis. In effect, this means that a formula of the form $!X$ can be read as "there is a reason supporting X" and also used to refer to the reason that grounds it. This is why we will often call such formulas *reasons*. We use $\mathcal{R}$ and $\mathcal{J}$ to denote (finite) collections of !-formulas: these will represent, respectively, *requiring* and *justifying reasons*—we adopt Gert's terminology. We also introduce the function $Conclusion(\cdot)$ that transforms !-formulas (and sets of such formulas) into ordinary propositional ones: thus, $Conclusion(!A) = A$ and $Conclusion(\{!L\}) = \{L\}$. The intuitive idea that some reasons have more weight than others will be captured by supplementing sets of !-formulas with a strict partial order. An expression of the form $!Y < !X$ should, then, be read as saying that the reason that grounds $!X$ has more weight than the reason that grounds $!Y$.[13]

We represent particular cases using the notion of a context:

Definition 3.1 [Contexts] A *context* Δ is a structure of the form $\langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$, where $\mathcal{W}$ is a consistent set of propositional formulas, $\mathcal{R}$ and $\mathcal{J}$ are finite sets of !-formulas, with the requirement that $\mathcal{R} \subseteq \mathcal{J}$, and $<$ is a strict partial order on $\mathcal{J}$.[14]

For illustration, we express Save One or Two in the context $\Delta_1 = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$ where $\mathcal{W} = \{(A\&B) \supset L\}$, $\mathcal{R} = \{!A, !B\}$, $\mathcal{J} = \{!A, !B, !\neg L\}$, and $<$ is empty.

[11] If your intuitions differ on this, consider upping the cost to yourself. Instead of losing your legs, you might lose your life.

[12] Similar notation is used in [17], [32, Appendix 2], and [45].

[13] We thank an anonymous reviewer for pressing us to clarify our conceptualization of formulas preceded by the ! (bang) operator.

[14] The constraint that $\mathcal{J}$ is finite keeps proofs in the appendix more manageable. It also fits well with the informal literature.

Given our intended interpretation of $\mathcal{R}$ and $\mathcal{J}$, the requirement that $\mathcal{R} \subseteq \mathcal{J}$ amounts to the idea that every requiring reason can serve as a justifying one. And the fact that $!\neg L$ is in $\mathcal{J}$, but not $\mathcal{R}$ formalizes the idea that the reasons that speaks against you losing your legs are purely justifying.

As our next step, we extend the language with three deontic operators. Thus, henceforth, we allow for formulas of the form $Ought(X)$, $Must(X)$, and $Can(X)$; and read them as saying, respectively, that it ought to be the case that X, that it is required, or that it must be the case, that X, and that it is permitted that X.[15] In what follows, we will often refer to these formulas as, respectively, *oughts*, *requirements*, and *permissions*. Before we specify a procedure for deriving such formulas from contexts, it pays noting that the emerging consensus in linguistics is that there are two distinct deontic necessities: a weaker one—typically ascribed using the modals *ought to* and *should*—and a stronger one—typically ascribed using *must* and *have to*.[16] Our oughts are meant to capture the weaker modality, while the requirements are meant to capture the stronger one.

Turning to the procedure, we need to specify how conflicts between reasons of different strength get resolved. A standard albeit simplistic move is to classify a reason r as "undefeated" if there is no stronger (requiring) reason r' such that $\mathcal{W} \cup Conclusion(r') \vdash \neg Conclusion(r)$.[17] Unfortunately, this approach won't do for us.[18] So, instead, we make use of a slightly more complex approach, motivated by the work of Brewka [4] and its characterization in [18, Ch. 8.2]. We start by defining two notions.

Definition 3.2 [Active reasons] Given a context $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$ and $\mathcal{D} \subseteq \mathcal{J}$, let
$Active_{\Delta}(\mathcal{D}) = \{r \in \mathcal{J} : \mathcal{W} \cup Conclusion(\mathcal{D}) \cup Conclusion(r) \nvdash \bot \text{ and } r \notin \mathcal{D}\}$.

Thus, a reason r is active relative to a set of reasons $\mathcal{D}$ in case it is consistent with $\mathcal{D}$, but not (yet) in $\mathcal{D}$. The second notion we need is that of $<$-maximal elements:

Definition 3.3 [Maximal element] Given a set of reasons $\mathcal{D}$ and a preorder $<$ on $\mathcal{D}$, let $Max_{<}(\mathcal{D}) = \{r \in \mathcal{D} : \text{there is no } r' \in \mathcal{D} \text{ with } r < r'\}$.

Here is the basic idea of the Brewka-motivated approach: given a context $\langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$, we look at all possible ways of extending $<$ to a total order $<'$ on $\mathcal{J}$, and then, for each of those ways, we build a set of reasons whose

[15] The distinction between impersonal and personal obligations—as well as requirements and permissions—is orthogonal to our goals. So, we follow what Horty [17] calls the *policy of intentional, but harmless equivocation* and move freely between impersonal and personal reading of $Ought(X)$, $Must(X)$, and $Can(X)$.

[16] For the discussion of linguistic data, see e.g., [46], [36, pp. 79–81]; for its importance for ethical theory and reasons-first views in particular, see [3], [42], and [43], and for its importance for deontic logic, see [24].

[17] See, e.g., [17], [20], [29].

[18] For a critical discussion of this approach and a number of others, see [18, Ch. 8]. We can't use it, because it gives rise to counterexamples to our Proposition 5.1.

conclusions are consistent, starting with the empty set and iteratively selecting the $<'$-maximal element from among the reasons that are active at a given step. Our next two definitions make this idea precise.

Definition 3.4 [Brewka scenarios, for totally ordered contexts] Let $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$ be a context where $<$ totally orders $\mathcal{J}$. Then $\mathcal{B}$ is the *Brewka scenario* of Δ just in case $\mathcal{B} = \bigcup_{i \geq 0} \mathcal{B}_i$, where the sequence $\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_2, \ldots$ is defined as follows:

$$\mathcal{B}_0 \quad = \emptyset,$$

$$\mathcal{B}_{i+1} = \begin{cases} \mathcal{D}_i & \text{if } \mathit{Active}_\Delta(\mathcal{B}_i) = \emptyset \\ \mathcal{D}_i \cup \mathit{Max}_<(\mathit{Active}_\Delta(\mathcal{B}_i)) & \text{otherwise} \end{cases}$$

To illustrate, consider the context $\Delta_2 = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$ where $\mathcal{W} = \{A \supset \neg B, B \supset \neg C\}$, $\mathcal{R} = \mathcal{J} = \{!A, !B, !C\}$ and $!A < !B < !C$. Notice that $!A$ and $!C$ are compatible, while $!B$ conflicts with both of them. Now let's determine the unique Brewka scenario $\mathcal{B}$ of this context by constructing the sequence $\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_2, \ldots$ such that $\mathcal{B} = \bigcup_{i \geq 0} \mathcal{B}_i$. Clearly, $\mathcal{B}_0$ is the empty set. Since $\mathit{Max}_<(\mathit{Active}_\Delta(\mathcal{B}_0))$ equals $\{!C\}$, we have $\mathcal{B}_1 = \{!C\}$. Further, it is not difficult to see that $\mathit{Max}_<(\mathit{Active}_\Delta(\mathcal{B}_1))$ equals $\{!A\}$. Since $\mathcal{W} \cup \mathit{Conclusion}(\mathcal{B}_1) = \{A \supset \neg B, B \supset \neg C, A\}$ entails $\neg B$, the reason $!B$ is not in $\mathit{Active}_\Delta(\mathcal{B}_1)$, while the reason $!C$ is. As a reasult, we have $\mathcal{B}_2 = \{!A, !C\}$. After this step, there are no further active reasons that could be added, and so we have $\mathcal{B}_i = \mathcal{B}_2$ for every $i \geq 2$. At this point it should be clear that the Brewka scenario $\mathcal{B} = \bigcup_{i \geq 0} \mathcal{B}_i$ that we were looking for is $\{!C, !A\}$.

Our next definition extends the notion of a Brewka scenario to contexts that are not totally ordered.

Definition 3.5 [Brewka scenarios] Let $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$ be any context. Then $\mathcal{B}$ is a *Brewka scenario* based on Δ just in case $\mathcal{B}$ is the Brewka scenario of some context $\langle \mathcal{W}, \mathcal{R}, \mathcal{J}, <' \rangle$ where $<'$ is a total order extending $<$.

Returning to the earlier Save One or Two scenario, there are six ways to extend the empty relation of Δ_1 to a total order. What is important to determining the Brewka scenarios of the resulting totally ordered contexts are only the two highest ranked reasons—how they are related to each other doesn't matter. If $!A$ and $!B$ are ranked the highest, the Brewka scenario is $\{!A, !B\}$. If $!A$ and $!\neg L$ are ranked the highest, the Brewka scenario is $\{!A, !\neg L\}$. Lastly, if $!B$ and $!\neg L$ are ranked the highest, we get $\{!B, !\neg L\}$. Thus, in total, there are three Brewka scenarios based on Δ_1.

In addition to Brewka scenarios, our procedure for deriving oughts, requirements, and permissions, will make use of the following auxiliary notion:

Definition 3.6 [Stable scenarios, restricted contexts] Given a context $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$, a *stable scenario* based on Δ is any set $\mathcal{D}$ such that $\mathcal{R} \subseteq \mathcal{D} \subseteq \mathcal{J}$. Letting $<^\mathcal{D}$ stand for $<$ restricted to $\mathcal{D}$, we call the context $\langle \mathcal{W}, \mathcal{R}, \mathcal{D}, <^\mathcal{D} \rangle$ the *restriction of* Δ *to* $\mathcal{D}$ and denote it by $\Delta^\mathcal{D}$.

So, a stable scenario includes all requiring reasons and any set of justifying ones—which implies that $\mathcal{R}$ always qualifies as a stable scenario. For illustration, there are two stable scenarios based on Δ_1: $\{!A, !B\}$ and $\{!A, !B, !\neg L\}$.

We are finally in a position to specify the conditions under which oughts, requirements, and permissions follow from contexts. We start with oughts. Intuitively, these are obtained by restricting attention to requiring reasons and completely ignoring the justifying ones, and then looking at what follows from all Brewka scenarios that can be constructed from them.

Definition 3.7 [Oughts] Given a context $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$, the formula $Ought(X)$ follows from Δ, written as $\Delta \mid\!\sim Ought(X)$, just in case, $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$ for every Brewka scenario $\mathcal{B}$ based on $\Delta^{\mathcal{R}}$.

It's not difficult to verify that $Ought(A \& B)$ follows from Δ_1: you ought to save both Alice and Bob. Whereas oughts are determined on the basis of requiring reasons alone, requirements and permissions are determined on the basis of both types of reasons. The idea underlying our definitions is simple: $Must(X)$ follows from a context when, for every stable scenario based on the context, X is a consequence of all of its Brewka scenarios; and $Can(X)$ follows when, for some stable scenario based on the context, X is a consequence of one of its Brewka scenarios.

Definition 3.8 [Requirements] Given a context $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$, the formula $Must(X)$ follows from it, $\Delta \mid\!\sim Must(X)$, just in case, for every stable scenario $\mathcal{D}$ based on Δ, we have $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$ for every Brewka scenario $\mathcal{B}$ based on $\Delta^{\mathcal{D}}$.

Definition 3.9 [Permissions] Given a context $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$, the formula $Can(X)$ follows from it, $\Delta \mid\!\sim Can(X)$, just in case, for some stable scenario $\mathcal{D}$ based on Δ, we have $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$ for some Brewka scenario $\mathcal{B}$ based on $\Delta^{\mathcal{D}}$.

For illustration, the two stable scenarios based on Δ_1, one of which we've discussed in detail above, give rise to three Brewka scenarios: $\{A, B\}$, $\{A, \neg L\}$, and $\{B, \neg L\}$. Since $A \vee B$ follows from all of them, we have $\Delta_1 \mid\!\sim Must(A \vee B)$. And since $A \vee B$ and $\neg L$ follow from some, we have $\Delta_1 \mid\!\sim Can(A \& B)$ and $\Delta_1 \mid\!\sim Can(\neg L)$. You *have to* save either Alice, or Bob; you can (and ought to) save both of them; and you can keep your legs. Thus, the model gets all the intuitions about Save One or Two right.

The model also has some nice properties. We register them here as a set of propositions—the proof of Proposition 3.2 is given in the appendix, the other two follow straightforwardly from the definitions:

Proposition 3.1 *For any context $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$, (i) if $\Delta \mid\!\sim Must(X)$, then $\Delta \mid\!\sim Ought(X)$; and (ii) if $\Delta \mid\!\sim Ought(X)$, then $\Delta \mid\!\sim Can(X)$.*

Proposition 3.2 *For any context $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$, neither $\Delta \mid\!\sim Ought(\bot)$, nor $\Delta \mid\!\sim Must(\bot)$, nor $\Delta \mid\!\sim Can(\bot)$.*

Proposition 3.3 *Let* $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$ *be an arbitrary context. Then (i)* $\Delta \mid\!\sim Ought(X \& Y)$ *just in case both* $\Delta \mid\!\sim Ought(X)$ *and* $\Delta \mid\!\sim Ougth(Y)$ *and (ii)* $\Delta \mid\!\sim Must(X \& Y)$ *just in case both* $\Delta \mid\!\sim Must(X)$ *and* $\Delta \mid\!\sim Must(Y)$*.*

Before we leave this section, let us answer two natural questions. The first concerns conditional oughts, requirements, and permissions. It's natural to wonder how these might be captured in our framework. It turns out that we can capture them by generalizing a familiar idea, going back at least to [16]. As a first step, we define the notion of updated contexts:

Definition 3.10 [Updated contexts] Given a context $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$ and a formula X consistent with $\mathcal{W}$, the result of updating, or supplementing, Δ with X, written as $\Delta[X]$, is the context $\langle \mathcal{W} \cup \{X\}, \mathcal{R}, \mathcal{J}, < \rangle$.

Thus, the context $\Delta[X]$ is just like Δ, except that X is now taken to be an established fact. With the notion of updated contexts, we can specify when conditional deontic statements follow from a context as follows:

Definition 3.11 [Conditional oughts, requirements, and permissions] Let Δ be an arbitrary context. Then:
$\Delta \mid\!\sim Ought(Y|X)$ just in case $\Delta[X] \mid\!\sim Ought(Y)$;
$\Delta \mid\!\sim Must(Y|X)$ just in case $\Delta[X] \mid\!\sim Must(Y)$;
$\Delta \mid\!\sim Can(Y|X)$ just in case $\Delta[X] \mid\!\sim Can(Y)$.

The second natural question concerns Definitions 3.7, 3.8, and 3.9: one may wonder what prompts the choice of what's known as the *disjunctive account* (over the *conflict account*).[19] The short answer is that not much seems to hinge on it, given our purposes, and that, in deontic settings, the disjunctive account is the less committal of the two and so also safer to work with.

4 Supererogation and the standard account

Having set up the formal model, let's take a step back from it and reconsider the Save One or Two scenario. As we have already noted, there seems to be an intuitive sense of *ought* in which you ought to save Alice *and* Bob, but it's not the case that you *have to* do it. Still, saving Alice and Bob is not only permissible, but would also be highly admirable. In fact, there seems to be a clear intuitive sense in which it is the best thing you could do. From a third-person perspective, it certainly looks like this action leads to the best possible outcome, with all three people involved staying alive—although one of them severely inured.

And this means that saving Alice and Bob is a *supererogatory action* as it is an action that is ostensibly best, and yet it isn't obligatory. What Muñoz [30] calls the *Classic Paradox of Supererogation* is the challenge to explain the very possibility of such actions. Our formal approach has the resources to meet this challenge—which it inherits from the core idea of dual-role approaches. Thus, in response to the question of why saving Alice and Bob is the best action,

[19] See, e.g., [17] for a discussion.

we can say that it maximizes compliance with the requiring reasons at play in the scenario. In fact, carrying out $A\&B$, that is, saving Alice and Bob, means complying with all the requiring reasons at play in Δ_1, that is, $!A$ and $!B$. And in response to the question of why saving Alice and Bob isn't obligatory, our approach lets us point to the (purely) justifying reason $!\neg L$ and say that it can serve as an excuse to not comply with one of the requiring reasons. Notice that these answers straightforwardly generalize to other cases involving supererogatory actions, giving us a general response to the classic paradox.

There's another formal approach to supererogation—the titular standard approach—that resolves the classic paradox, namely, McNamara's *Doing Well Enough* framework [23], [24], [25].[20] McNamara works with ranked possible worlds: the higher a world's ranking, the (morally) better it is. Requirements are determined by a threshold: if X is true in all worlds above it, it's required that X. Permissions are duals of requirements: if it's not the case that $\neg X$ is required, it's permissible that X. Oughts in our sense are determined by the best worlds, they are "the most one can do": if X is true in all the best worlds, it ought to be that X. Also, since the best worlds are above the threshold, this gives the intuitive principle that requirements imply oughts.

This setup lets McNamara account for the intuitions in Save One or Two and respond to the challenge: saving Alice and Bob is best because the worlds where both get saved are ranked the highest; it is not obligatory because there are other worlds above the threshold where only one person is saved.[21]

So now we have seen two formal accounts of supererogation. The standard one might look more elegant and simple, but there's a serious problem with an account like this: transitively ranking all worlds and determining acceptability by means of a threshold imposes serious restrictions. It rules out scenarios where an impermissible act is superior to a permissible one—cf. [47]. The problem is that such scenarios seem possible:[22]

> **All or Nothing**. Alice and Bob are, again, trapped in a collapsing building, but this time you will lose your legs whether you save one or both of them.[23]

Intuitively, worlds where only one person is saved are superior to those where none are. Nevertheless, walking away seems permissible, while saving only one person does not—it involves gratuitous loss of life. The threshold framework

[20] We see McNamara's work as a representative of the dominant approach to deontic modality in philosophy and linguistics, associated, among others, with [21] and [22]. The difficulties that McNamara faces are symptomatic of problems for this dominant approach. This is evidenced by the fact that Åqvist in [2], who defends an even more fine-grained threshold model with an arbitrary number of levels of goodness, still cannot accommodate the scenario we discuss in the next section in a natural and intuitive manner—at least not without giving up the intuitive notion of a threshold.

[21] Perhaps, a fully satisfactory explanation would need to say more about the ranking and threshold, but there are several plausible things to say here.

[22] In Section 6, we consider the question of how the standard account might be changed to address this problem.

[23] The case comes from [14].

says otherwise: since worlds where one person gets saved are better than those where none are, and it's permissible to walk away, it must be permissible for you to save only one.

Our model, by contrasts, easily handles the case. We express it as the context $\Delta_3 = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$, where $\mathcal{W} = \{A \supset L, B \supset L\}$, $\mathcal{R} = \{!A, !B\}$, $\mathcal{J} = \{!A, !B, !\neg L\}$, and $<$ is empty. It's not difficult to verify that $Can(\neg A \& \neg B)$ follows from Δ_3, and that neither $Can(A \& \neg B)$, nor $Can(\neg A \& B)$ do: while it's permissible for you to save neither Alice, nor Bob, it's not the case that you can save only one.

5 Horton's All or Nothing Problem

In addition to the classic paradox, supererogation gives rise to at least two other puzzles. Following [30], we call them the *All or Nothing Problem* [14] and the *Intransitivity Paradox* [19]. While our dual-role approach can resolve both puzzles, here we discuss only the former one, for reasons of space. It emerges as combinations of intuitions about the All or Nothing scenario and a plausible principle. We present the problem as a set of jointly inconsistent claims in English, staying close to Horton's [14] original formulation:

1. It's morally permissible to save neither Alice nor Bob. (intuition)
2. It's morally wrong for you to save only one of them. (intuition)
3. If an act X is morally permissible and an act Y is morally wrong—and X and Y are the only two available acts—one ought to do X, rather than Y. (intuitive principle)
4. You ought to save neither Alice nor Bob rather than save only one of them. (from 1–3)
5. But, clearly, (4) is false. (intuition)

Two notes are in order. First, the *oughts* in claims (3) and (4) aren't meant to immediately map onto our technical notion of ought. Rather, at this point, claims (1)–(5) are meant to express pretheoretical intuitions—as they do in Horton's statement of the puzzle. Second, the paradox appeals to the notion of comparative obligations. While this notion makes intuitive sense and is used by Horton, it certainly hasn't been the focus of much research in deontic logic. Luckily, it seems possible to capture this notion in terms of conditional obligations: to say that one ought to do X, rather than Y is just to say that one ought to do X in case $X \vee Y$.[24] Bearing this in mind and letting A and B express the same propositions they did before, we propose to express the problem in our formal notation as follows—which, we contend, sharpens it:

1. $Can(\neg A \& \neg B)$ (intuition)
2. $Must(\neg([A \& \neg B] \vee [\neg A \& B]))$ (intuition)
3. If $Can(X)$ and $Must(\neg Y)$, then $Must(X | X \vee Y)$ (intuitive principle)

[24] In his original statement of the problem, Horton suggests this much—see [14, fn. 2].

4. If $Can(\neg A\&\neg B)$ and $Must(\neg([A\&\neg B] \vee [\neg A\&B]))$, then $Must(\neg A\&\neg B|[\neg A\&\neg B] \vee ([A\&\neg B] \vee [\neg A\&B]))$ (instance of the principle)
5. $Must(\neg A\&\neg B|[\neg A\&\neg B] \vee ([A\&\neg B] \vee [\neg A\&B]))$ (from 1, 2, and 4)
6. $Must(\neg A\&\neg B|\neg[A\&B])$ (substitution of equivalent formulas)
7. But, clearly, not $Must(\neg A\&\neg B|\neg[A\&B])$ (intuition)

It's worth being explicit about two assumptions in the background of our formalization. First, we take the oughts in the original claims to express the stronger deontic modals, what we called *requirements.* Second, we are assuming that if an action is morally wrong, there's a requirement forbidding taking this action. Both assumptions strike us as very plausible. What our formalization, then, does is show that All or Nothing is, indeed, a genuine puzzle, and that, their intuitive character notwithstanding, we cannot hold onto claims (1)–(3) and (7) on pain of inconsistency.

Our model happens to solve this puzzle, suggesting that the fault lies with the principle expressed in (3). First off, the principle's counterpart

If $\Delta \mid\!\sim Can(X)$ *and* $\Delta \mid\!\sim Must(\neg Y)$, *then* $\Delta \mid\!\sim Must(X|X \vee Y)$

is demonstrably false. This is witnessed by the context $\Delta_4 = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$ where $\mathcal{W} = \{C \supset \neg D, D \supset \neg E, E \supset \neg C\}$, $\mathcal{R} = \{!C, !D\}$, $\mathcal{J} = \{!C, !D, !E\}$, and $!D < !C$. It's quite easy to verify that we have both $\Delta_4 \mid\!\sim Can(E)$ and $\Delta_4 \mid\!\sim Must(\neg D)$, while we don't have $\Delta_4 \mid\!\sim Must(E|D \vee E)$. What's more, it can be shown that two principles in the vicinity hold true in the model—the proofs are provided in the appendix:

Proposition 5.1 *For any context* Δ,
(i) if $\Delta \mid\!\sim Can(X)$ *and* $\Delta \mid\!\sim Must(\neg Y)$, *then* $\Delta \mid\!\sim Can(X|X \vee Y)$;
(ii) if $\Delta \mid\!\sim Ought(X)$ *and* $\Delta \mid\!\sim Must(\neg Y)$, *then* $\Delta \mid\!\sim Ought(X|X \vee Y)$.[25]

The fact that these principles hold can explain the intuitive pull of the original principle. Our approach also makes clear where the original principle goes wrong: it attempts to bridge unconditional and conditional deontic statements without keeping track of the types of reasons that these statements depend on.

6 Related work

This section compares our model to Mullin's [29] dual-role approach to reasons and briefly discusses related work on supererogation.[26] After discussing

[25] To be fair, both principles are immediate consequences of more general principles that hold in the model, as the proofs in the appendix make manifest. An anonymous reviewer suggests that this weakens our claim that the principles we propose account for the intuitive pull of the original principle. While we share the intuition that, it would be a nice feature of the model, if our principles wouldn't be mere corollaries of more general ones, it is not immediately clear to us why the claim is weakened. In any event, Proposition 5.1 is the best we have for now, and it might well be that our model validates other principles that could serve its function, or serve it better.

[26] We focus on recent work on supererogation. For a historical perspective and its relevance to current topics see [26].

Mullins, we revisit McNamara's threshold account and consider his recent extension to conditional operators. Then we discuss Wessels' [47] quasi decision-theoretic approach—which discusses cases like All or Nothing—and Hansson's [13] approach.

6.1 Mullins dual-role approach

Mullins starts with Horty's model [18] and discusses two ways to capture the distinction between requirements and oughts. The first appeals to a threshold of strength, the second one—which we focus on—distinguishes between two distinct types of reasons in the spirit of dual-role approaches. [27]

Unlike us, Mullins relies on the simple approach to defeat, and his strategy is to, first, specify when a context entails a requirement—that is, a *Must*-formula—and then, in the second step, use this as a basis for determining which permissions and oughts this context entails. More precisely, $Can(X)$ is set to follow from a context just in case $Must(\neg X)$ does not follow, and $Ought(X)$ is set to follow just in case, roughly, the reasons that entail X are compatible with the reasons that allow for the derivation of *Must*-formulas. Explaining his strategy, Mullins writes:

> We first identify our undefeated requiring reasons, in order to determine what is required or impermissible. Oughts are then supported by our best justifying reasons, provided the consequences of their conclusions are consistent with some maximal subset of requiring reasons [29, p. 586].

To see Mullins' model at work, we revisit the familiar Save One or Two scenario. We captured it in the context $\Delta_1 = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$ where $\mathcal{W} = \{(A\&B) \supset L\}$, $\mathcal{R} = \{!A, !B\}$, $\mathcal{J} = \{!A, !B, !\neg L\}$, and $<$ empty. To determine which *Must*-formulas follow from it, we are to look at what follows from the subsets of undefeated requiring reasons—that is, the subsets of $Conclusion(\mathcal{R}) = \{A, B\}$—that are maxiconsistent with $\mathcal{W}$. This, however, gives us the counterintuitive result that both $Must(A\&B)$ and $Must(L)$ follow from Δ_1: you have to save Alice and Bob, and lose your legs. One might take this to mean that the scenario has to be captured in a different context, and the most natural alternative that suggests itself is $\Delta_5 = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$ where $\mathcal{W} = \{(A\&B) \supset L\}$, $\mathcal{R} = \emptyset$, $\mathcal{J} = \{!A, !B, !\neg L\}$, and $<$ is empty. [28] Even barring the counterintuitive im-

[27] See [29, Secs. 4 and 5]. The basic idea behind the first way to capture the distinction is that only reasons above a certain threshold can support requirements. Mullins attributes the idea to Scanlon [39].

[28] Another possible candidate is the context $\Delta_6 = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$ where $\mathcal{R} = \{!(A \vee B)\}$ and the rest is like in Δ_5. But while this secures the intuitive results that both $Must(A \vee B)$ and $Ought(A\&B)$ are derivable, there's good reason to be dissatisfied with this context. Most importantly, the inclusion of a *disjunctive* requiring reason looks terribly ad hoc, since it amounts to hard-coding the desired intuition. Also, $Ought(\neg L)$ follows from Δ_6, just like it does in the case of Δ_5. An anonymous reviewer worries that leaving the ordering $<$ empty in the representation of the case stacks the cards, since, in Mullins' model, requiring reasons can get defeated by justifying ones. While the reviewer's reaction is certainly reasonable, we couldn't think of any way the ordering might be used to get Mullins' account to deliver the right result: setting $!A$, $!B < !\neg L$ fails to deliver the intuitive $Must(A \vee B)$, while adding

plication that the fact that Alice and Bob are in danger doesn't exert requiring normative force, we don't get a good match with intuitions. First, $Must(A \vee B)$ doesn't follow from Δ_5. Second, while $Ought(A\&B)$ follows from Δ_5, so does $Ought(\neg L)$, suggesting that not losing your legs is optimal.[29]

This invites the conclusion that Mullins' model has serious trouble accommodating the Save One or Two scenario. There are other issues with it too, but we won't dwell on them and simply state (what we take to be) the underlying problem: it determines requirements almost exclusively on the basis of requiring reasons, not giving justifying reasons their due.[30] Admittedly, this appears to be the default approach in the (informal) philosophical literature—see, e.g., [43]—but it doesn't seem to work once expressed in a defeasible logic-based framework.

6.2 Doing Well Enough

Since McNamara's framework was already introduced in Section 4, here we confine ourselves to some brief remarks focusing on its conditionalized version, as developed in [27], and briefly sketch some worries whether, if at all, it might accommodate cases like Horton's *All or Nothing* scenario. The main advance of [27] is the provision of formal tools to capture *conditionally* acceptable worlds, in the style of Dyadic Deontic Logic [33, Ch. 2]. Acceptable worlds are those worlds that are above the threshold or "good enough", and any proposition true in one of these world is permitted. This allows one to formalize Horton's All or Nothing Problem using conditional obligations, like we suggest in Section 5. McNamara's analogues of our *Must*-, *Can*-, and *Ought*-operators are, respectively, $OB(\cdot)$, $PE(\cdot)$, and $MA(\cdot)$, "the most one can do". Both $OB(\cdot)$ and $MA(\cdot)$ function like standard dyadic operators. This has the consequence that the principle $PE(X)\&OB(\neg Y) \supset OB(X|X \vee Y)$ is a theorem in McNamara's logic as he presents it in [27]. The fact that this principle holds, depends crucially on the semantic principle that there is a threshold: if a world is acceptable, then any world better than it is acceptable as well. An anonymous reviewer notes that it is possible to give up this principle. This is true, but a challenge remains: one has to account for the (remaining) claims that comprise the All or Nothing puzzle without simply hard-coding which actions are

$!(A\&B)$ to $\mathcal{R}$ and setting $!(A\&B) < !\neg L$ (as the reviewer appears to suggest) doesn't really change anything.

[29] Mullins' strategy uses the *conflict account* in determining which oughts follow from a context. A natural idea is to substitute it with the *disjunctive account*—Shyam Nair suggested this much in his keynote talk at the DEON2020/21 conference. Unfortunately, this move doesn't solve the problem: $Ought(\neg L)$ no longer follows from the context, but neither does $Ought(A\&B)$.

[30] There is only one way in which justifying reasons can have an impact on the requirements in Mullins' model: they can outright defeat requiring reasons. This, however, appears to be not enough to get the cases right—see Footnote 28. We thank an anonymous reviewer for pushing us to clarify our take on Mullins' approach.

permitted and which are not into the logical description of the case. [31]

All in all, while McNamara's (conditional) system allows one to express normative notions that we cannot easily capture—like "the least one can do"—we take it to be a serious challenge to modify it so that it can account for puzzles surrounding supererogation (of which All or Nothing is one) in a natural way. [32] In any event, we should be wary of any framework committed to the existence of a threshold, since it implies some of our intuitions about the All or Nothing scenario must be mistaken.

6.3 Other recent proposals

Wessels [47] proposes a very different account to accommodate supererogation. It is not a full-fledged logic, but still instructive. Wessels' explicit goal is to account for cases like the All or Nothing scenario, or cases involving what she calls "supererogation holes".

In Wessels' account, actions (instead of worlds) are totally ordered by their respective goodness, and an actions' "being supererogatory with respect to another action" is used to define supererogation simpliciter. The core idea is that an action is supererogatory with respect to another action just in case the relation between gained moral value and burden to the agent is above a threshold. [33] Using this construction, Wessels then defines supererogatory actions as follows.

> An action f_j is supererogatory just in case the answers to all three subquestions is yes:
> (1) Is there an action f_i such that f_j is supererogatory with respect to f_i?
> (2) Are all the actions that are morally better than f_j supererogatory with respect to f_j?
> (3) Are all the actions that are morally better than f_i supererogatory with respect to f_i?

Notice how this way of capturing supererogation lets her say that, in the All or Nothing scenario, saving either only Alice or only Bob is not supererogatory: since the act of saving both Alice and Bob doesn't put an additional burden on the agent when compared to saving either only Alice or only Bob, it's not supererogatory with respect to these other acts, and so the answer to the second question is negative.

[31] Note that our account makes no such assumptions. Which (conditional) actions are permitted, obligatory, or required follows from the interplay of reasons and their strength alone.

[32] In addition to the issue discussed in the previous paragraph, McNamara's framework faces a second problem, which we can only hint at here. As is, it validates the principle $PE(X|Y \vee X) \& PE(Z|X \vee Y) \supset PE(Z|Y \vee Z)$ which certain well-known cases involving supererogation bring into doubt. Here, too, the framework would have to be modified to account for this fact. See [19], as well as [31] for a discussion.

[33] Wessels uses real numbers to represent this in the style of rational choice theory with some restrictions. For instance, one action is allowed to be supererogatory with respect to another one just in case the moral value of the first action is at least as high as the moral value of the second one.

While we set out to develop a formal model that can solve some puzzles surrounding supererogation, Wessels' aims are more moderate. What her framework shows, in effect, is that if a logic can solve the problem of defining "X supererogatory with respect to Y", then a preference-based logic might take care of the rest. What Wessels doesn't do is develop such a logic.[34]

Hansson [13], unlike Wessels, proposes a logic-based account of supererogation which, like Wessels' account, builds on the relation "p is supererogatory relative to q". Hansson's idea is to set it that p is supererogatory relative to q if q is obligatory and p is "a better variant of q". Betterness is spelled out in terms of a preference relation, whereas "is a version of" is a primitive spelled out in terms of logical strength: thus, $p \vdash q$ means that p is a variant of q. This approach seems to be overly simplistic as it faces two challenges. First, not every supererogatory action is a variant of some obligatory action. For example, in the All or Nothing case, the supererogatory action is saving Alice and Bob, while no action at all appears to be obligatory. The second challenge is that, in modeling such scenarios, the choice of which actions are variants of one another is threatened to become entirely ad hoc.[35]

7 Conclusion

We set ourselves two goals in this paper. The first was to express the core of dual-role approaches to reasons in a defeasible logic. To reach this goal, we extended Horty's influential default logic-based model [18] in a number of ways. Our second goal had to do with supererogation, and we saw how our dual-role approach provides a unified response to the Classic Paradox of Supererogation and the All or Nothing Problem. What's more, we noted some advantages that our model has over alternative (formal) approaches to supererogation.

We see several promising directions for future research. First, our approach seems to let us solve another notorious puzzle about supererogation, namely, Kamm's *Intransitivity Paradox* [19], and we plan to discuss the issue in detail in a follow-up paper. Second, it would be interesting to explore how dual-role approaches to reasons might be captured in other frameworks that have been used to model reasons, such as structured argumentation or justification logic. Relatedly, it seems worthwhile to relate our model to input/output logic with permission—the latter looks like a more general system. Third, it might pay exploring further applications of formalized dual-role approaches. For instance, our model might have something interesting to say about the puzzles associated with permission. Lastly, it's worth thinking about the commitments of the particular dual-role approach that comes with the model and its potential advantages over the dual-role views defended in the philosophical literature.

[34] See [26] for a critical discussion of Wessels in the context of deontic logic.

[35] It pays noting that, in a critical discussion of Hansson [13], McNamara [26] suggests a way to ameliorate the second challenge by means of introducing a dyadic action operator standing for "an agent brings it about that q by bringing about that p". We suspect, however, that this move makes the first challenge more pressing, since it imposes further restrictions on what can count as a variant of an action.

Appendix

Proposition 3.2 *For any context $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, < \rangle$, neither $\Delta \mathrel{|\!\sim} Ought(\bot)$, nor $\Delta \mathrel{|\!\sim} Must(\bot)$, nor $\Delta \mathrel{|\!\sim} Can(\bot)$.*

Proof. It is enough to show that $\Delta \mathrel{|\!\sim} Can(\bot)$ does not hold, since by Proposition 3.1, if one of the other statements were to hold, so would $\Delta \mathrel{|\!\sim} Can(\bot)$.

Suppose, toward a contradiction, that $\Delta \mathrel{|\!\sim} Can(\bot)$. This implies that there is a stable scenario $\mathcal{D}$ and a Brewka scenario $\mathcal{B}$ based on $\Delta^{\mathcal{D}} = \langle \mathcal{W}, \mathcal{R}, \mathcal{D}, <^{\mathcal{D}} \rangle$ such that $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash \bot$. Let $<^*$ be the ordering that extends $<^{\mathcal{D}}$ to a total order over $\mathcal{D}$, and that is used in the construction of $\mathcal{B}$. Given that $\mathcal{W}$ is consistent and that $\mathcal{B}$ is the limit of the sequence $\mathcal{B}_0, \mathcal{B}_1, \ldots$, we can be sure that there is some i such that $\mathcal{W} \cup Conclusion(\mathcal{B}_i) \nvdash \bot$, while $\mathcal{W} \cup Conclusion(\mathcal{B}_{i+1}) \vdash \bot$. But $\mathcal{B}_{i+1} = \mathcal{B}_i \cup Max_{<^*}(Active_{\langle \mathcal{W}, \mathcal{R}, \mathcal{D}, <^* \rangle}(\mathcal{B}_i))$, and $Max_{<^*}(Active_{\langle \mathcal{W}, \mathcal{R}, \mathcal{D}, <^* \rangle}(\mathcal{B}_i))$ is the singleton set $\{r \in \mathcal{D} : \mathcal{W} \cup Conclusion(\mathcal{B}_i) \cup Conclusion(r) \nvdash \bot \text{ and } r \notin \mathcal{B}_i\}$. Given that $\mathcal{W} \cup Conclusion(\mathcal{B}_i) \cup Conclusion(r) = \mathcal{W} \cup Conclusion(\mathcal{B}_{i+1})$, we have arrived at a contradiction. □

Before we turn to the proof of Proposition 5.1, we establish two lemmas.

Lemma 1 Given a context $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, <' \rangle$, where $<'$ is a total order over $\mathcal{J}$, and a Brewka scenario $\mathcal{B}$ based on Δ, we have $\mathcal{W} \cup Cocnlusion(\mathcal{B}) \nvdash \bot$.

Proof. The lemma follows by an easy induction on the construction of $\mathcal{B}$. □

Lemma 2 Let $\Delta = \langle \mathcal{W}, \mathcal{R}, \mathcal{J}, <' \rangle$ be a context with $<'$ a total order on $\mathcal{J}$, $\mathcal{B}$ the Brewka scenario of Δ with $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$, and $\mathcal{B}^*$ the Brewka scenario of the context $\Delta[X \vee Y] = \langle \mathcal{W} \cup \{X \vee Y\}, \mathcal{R}, \mathcal{J}, <' \rangle$. Then $\mathcal{B} = \mathcal{B}^*$.

Proof. Before getting into the proof, note that both $\mathcal{B}$ and the sequence it is the limit of are unique. This is due ot the fact that at each step i of the construction of $\mathcal{B}$ there's at most one $<'$-maximal reason in $Active_{\Delta}(\mathcal{B}_i)$. We will show that, for each step i, $\mathcal{B}_i = \mathcal{B}_i^*$. We do this by induction on i.

The base case is trivial: $\mathcal{B}_0 = \emptyset = \mathcal{B}_0^*$.

For the induction step, assume that $\mathcal{B}_i = \mathcal{B}_i^*$. Given our definition of Brewka scenarios, it's enough to establish that $Max_{<'}(Active_{\Delta}(\mathcal{B}_i)) = Max_{<'}(Active_{\Delta[X \vee Y]}(\mathcal{B}_i^*))$. So that's what we turn to.

$\subseteq$: Consider some $r \in Max_{<'}(Active_{\Delta}(\mathcal{B}_i))$. Then $r \in \mathcal{B}_{i+1} \subseteq \mathcal{B}$. We know that $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$, and hence that $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X \vee Y$. By Lemma 1, $\mathcal{W} \cup Conclusion(\mathcal{B})$ is consistent. Hence, as it entails $X \vee Y$, it is also consistent with $X \vee Y$. Since $\mathcal{W} \cup Conclusion(r) \cup Conclusion(\mathcal{B}_i)$ is a subset of $\mathcal{W} \cup Conclusin(\mathcal{B})$, we can be sure that $\mathcal{W} \cup \{X \vee Y\} \cup Conclusion(B_i) \cup Conclusion(r) \not\vdash \bot$. This suffices to conclude that $r \in Active_{\Delta[X \vee Y]}(\mathcal{B}_i)$. We still need to show that r is $<'$-maximal in this set. We know that r is $<'$-maximal in $Active_{\Delta}(\mathcal{B}_i)$. Hence, for every $r' > r$ such that $r' \notin \mathcal{B}_i$, $r' \notin Active_{\Delta}(\mathcal{B}_i)$. But this means that, for every such r', $\mathcal{W} \cup Conclusion(\mathcal{B}_i) \cup Conclusion(r') \vdash \bot$, and, hence, by monotonicity, that $\mathcal{W} \cup \{X \vee Y\} \cup Conclusion(\mathcal{B}_i) \cup Conclusion(r') \vdash \bot$. This means that each such r' is not in $\mathit{Active}_{\Delta[X\vee Y]}(\mathcal{B}_i)$, and that r is indeed maximal here. By the inductive hypothesis, $\mathcal{B}_i = \mathcal{B}_i^*$, and so we have shown that $r \in \mathit{Max}_{<'}(\mathit{Active}_{\Delta[X\vee Y]}(\mathcal{B}_i^*))$.

$\supseteq$: Suppose that there is an $r \in \mathit{Max}_{<'}(\mathit{Active}_{\Delta[X\vee Y]}(\mathcal{B}_i^*))$. Further, suppose, toward a contradiction, that $r \notin \mathit{Max}_{<'}(\mathit{Active}_{\Delta}(\mathcal{B}_i))$. Either $r \in \mathit{Active}_{\Delta}(\mathcal{B}_i)$ or not. If not, then either (i) $r \in \mathcal{B}_i$, or (ii) $\mathcal{W} \cup \mathit{Conclusion}(\mathcal{B}_i) \cup \mathit{Conclusion}(r) \vdash \bot$. If (i), then $\mathcal{B}_i \neq \mathcal{B}_i^*$. If (ii), then $\mathcal{W} \cup \mathit{Conclusion}(\mathcal{B}_i^*) \cup \mathit{Conclusion}(r) \vdash \bot$, and $r \notin \mathit{Active}_{\Delta[X\vee Y]}(\mathcal{B}_i^*)$ after all. Hence, $r \in \mathit{Active}_{\Delta}(\mathcal{B}_i)$, but not $<'$-maximal. Let r' be the $<'$-maximal reason in $\mathit{Active}_{\Delta}(\mathcal{B}_i)$. Now we can reuse the argument we made use of above to conclude that $r' \in \mathit{Active}_{\Delta[X\vee Y]}(\mathcal{B}_i^*)$, and that $r \notin \mathit{Max}_{<'}(\mathit{Active}_{\Delta[X\vee Y]}(\mathcal{B}_i^*))$ after all. This gives us a contradiction. □

Proposition 5.1 *For any context* Δ,

(i) if $\Delta \mathrel{|\!\sim} \mathit{Can}(X)$ *and* $\Delta \mathrel{|\!\sim} \mathit{Must}(\neg Y)$*, then* $\Delta \mathrel{|\!\sim} \mathit{Can}(X|X \vee Y)$*;*

(ii) if $\Delta \mathrel{|\!\sim} \mathit{Ought}(X)$ *and* $\Delta \mathrel{|\!\sim} \mathit{Must}(\neg Y)$*, then* $\Delta \mathrel{|\!\sim} \mathit{Ought}(X|X \vee Y)$*.*[36]

Proof. We establish claim (i) by proving a stronger claim, namely, that if $\Delta \mathrel{|\!\sim} \mathit{Can}(X)$, then $\Delta \mathrel{|\!\sim} \mathit{Can}(X|X \vee Y)$. Suppose that $\Delta \mathrel{|\!\sim} \mathit{Can}(X)$. It follows that there is a stable scenario $\mathcal{D}$ based on Δ and a total order $<'$ on $\mathcal{D}$ that extends $<$ such that $\mathcal{W} \cup \mathit{Conclusion}(\mathcal{B}) \vdash X$ for the Brewka scenario based on $\langle \mathcal{W}, \mathcal{R}, \mathcal{D}, <' \rangle$. Note that $\mathcal{D}$ is a stable scenario of $\Delta[X \vee Y]$, and that $<'$ is a total order extending $<$ in this restricted updated context as well. Set Δ^* to be the context $\langle \mathcal{W} \cup \{X \vee Y\}, \mathcal{R}, \mathcal{D}, <' \rangle$. By Lemma 2, we know that $\mathcal{B}^* = \mathcal{B}$ where $\mathcal{B}^*$ is the Brewka scenario based on Δ^*. Since $\mathcal{W} \cup \mathit{Conclusion}(\mathcal{B}) \vdash X$, we immediately get $\mathcal{W} \cup \mathit{Conclusion}(\mathcal{B}^*) \vdash X$, and, by monotonicity of classical logic, $\mathcal{W} \cup \{X \vee Y\} \cup \mathit{Conclusion}(\mathcal{B}^*) \vdash X$. This means that there is a stable scenario of $\Delta[X \vee Y]$, namely, $\mathcal{D}$, and a Brewka scenario based on $\Delta[X \vee Y]^{\mathcal{D}}$, namely, $\mathcal{B}^*$, such that $\mathcal{W} \cup \{X \vee Y\} \cup \mathit{Conclusion}(\mathcal{B}^*) \vdash X$. Given our definition of permissions, this is enough to conclude that $\Delta[X \vee Y] \mathrel{|\!\sim} \mathit{Can}(X)$, and hence that $\Delta \mathrel{|\!\sim} \mathit{Can}(X|X \vee Y)$.

For Claim (ii), we prove something stronger, namely, that if $\Delta \mathrel{|\!\sim} \mathit{Ought}(X)$, then also $\Delta \mathrel{|\!\sim} \mathit{Ought}(X|X \vee Y)$. Suppose that $\Delta \mathrel{|\!\sim} \mathit{Ought}(X)$. This means that, for any Brewka scenario $\mathcal{B}$ of the context $\Delta^{\mathcal{R}}$, we have $\mathcal{W} \cup \mathit{Conclusion}(\mathcal{B}) \vdash X$. What we need to show is that, for any Brewka scenario $\mathcal{B}^*$ based on $\Delta[X \vee Y]^{\mathcal{R}}$, we have $\mathcal{W} \cup \{X \vee Y\} \cup \mathit{Conclusion}(\mathcal{B}^*) \vdash X$. Suppose, toward a contradiction, that this wasn't the case. So there is a context $\Delta^* = \langle \mathcal{W} \cup \{X \vee Y\}, \mathcal{R}, \mathcal{R}, <' \rangle$, where $<'$ extends $<$ to a total order over $\mathcal{R}$, such that $\mathcal{W} \cup \{X \vee Y\} \cup \mathit{Conclusion}(\mathcal{B}^*) \nvdash X$ for the

[36] To be fair, both principles are immediate consequences of more general principles that hold in the model, as the proofs in the appendix make manifest. An anonymous reviewer suggests that this weakens our claim that the principles we propose account for the intuitive pull of the original principle. While we share the intuition that, it would be a nice feature of the model, if our principles wouldn't be mere corollaries of more general ones, it is not immediately clear to us why the claim is weakened. In any event, Proposition 5.1 is the best we have for now, and it might well be that our model validates other principles that could serve its function, or serve it better.

Brewka scenario $\mathcal{B}^*$ based on Δ^*. Consider the context $\langle \mathcal{W}, \mathcal{R}, \mathcal{R}, <' \rangle$. From above, we can be sure that, for the Brewka scenario $\mathcal{B}$ based on it, we have $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$. By Lemma 2, we have $\mathcal{B}^* = \mathcal{B}$. (Recall that $\mathcal{B}^*$ is unique.) Hence, $\mathcal{W} \cup Conclusion(\mathcal{B}^*) \vdash X$, and, by the monotonicity of classical logic, $\mathcal{W} \cup \{X \vee Y\} \cup Conclsuion(\mathcal{B}^*) \vdash X$. And this is a contradiction. □

References

[1] Alvarez, M., *Reasons for action: Justification, motivation, explanation*, in: E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, 2016, winter 2016 edition .

[2] Åqvist, L., *Three characterizability problems in deontic logic*, Nordic Journal of Philosophical Logic **5** (2000), pp. 65–82.

[3] Bedke, M., *Passing the deontic buck*, Oxford Studies in Metaethcis **6** (2011), pp. 128–53.

[4] Brewka, G., *Reasoning about priorities in default logic*, in: *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, 1994, pp. 940–5.

[5] Dancy, J., *Enticing reasons*, in: R. J. Wallace, P. Pettit, S. Scheffler and M. Smith, editors, *Reasons and Values: Themes from the Moral Philosophy of Joseph Raz*, Oxford University Press, 2004 pp. 91–118.

[6] Dietrich, F. and C. List, *A reason-based theory of rational choice*, Noûs **47(1)** (2013), pp. 104–34.

[7] Faroldi, F., *Common law precedent in the logic of reasons*, in: S. Rahman, M. Armgardt and H. Kvernenes, editors, *New Systems and Historic Studies in Legal Reasoning and Logic*, Springer, 2022 pp. 301–19.

[8] Gert, J., "Brute Rationality: Normativity and Human Action," Oxford University Press, 2004.

[9] Gert, J., *Normative strength and the balance of reasons*, Philosophical Review **116** (2007), pp. 533–62.

[10] Gert, J., *Perform a justified option*, Utilitas **26** (2014), pp. 206–17.

[11] Greenspan, P., *Asymmetrical practical reasons*, in: J. C. Marek and M. E. Reicher, editors, *Experience and Analysis: Proceedings of the 27th International Wittgenstein Symposium*, Vienna: ÖBV and HPT, 2005 pp. 387–94.

[12] Greenspan, P., *Practical reasons and moral 'ought'*, Oxford Studies in Metaethcis **2** (2007), pp. 181–205.

[13] Hansson, S. O., *Representing supererogation*, Journal of Logic and Computation **25** (2015), pp. 443–451.

[14] Horton, J., *The all or nothing problem*, Journal of Philosophy **114** (2017), pp. 94–104.

[15] Horty, J., "The Logic of Precedent: Constraint and Freedom in Common Law Reasoning," Cambridge University Press, forthcoming.

[16] Horty, J. F., *Moral dilemmas and nonmonotonic logic*, Journal of Philosophical Logic **23** (1994), pp. 35–65.

[17] Horty, J. F., *Reasoning with moral conflicts*, Noûs **37** (2003), pp. 557–605.

[18] Horty, J. F., "Reasons as Defaults," New York: Oxford University Press, 2012.

[19] Kamm, F., *Supererogation and obligation*, Journal of Philosophy **82** (1985), pp. 118–38.

[20] Knoks, A., *Misleading higher-order evidence, conflicting ideals, and defeasible logic*, Ergo **8** (2021), pp. 141–74.

[21] Kratzer, A., "Modals and Conditionals: New and Revised Perspectives," Oxford: Oxford University Press, 2012.

[22] Lewis, D., "Counterfactuals," Blackwell, 1973.

[23] McNamara, P., *Doing well enough: Toward a logic for common-sense morality*, Studia Logica **57** (1996), pp. 167–92.

[24] McNamara, P., *Must I do what I ought (or will the least I can do do)?*, in: M. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems*, Springer-Verlag, 1996 pp. 154–73.

[25] McNamara, P., *Praise, blame, obligation, and DWE: Toward a framework for classical supererogation and kin*, Journal of Applied Logic **9** (2011), pp. 153–70.

[26] McNamara, P., *Logics for supererogation and allied normative concepts*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, College Publications, 2022 pp. 155–306.

[27] McNamara, P., *A natural conditionalization of the dwe framework*, in: *Agency, Norms, Inquiry, and Artifacts: Essays in Honor of Risto Hilpinen*, Springer, 2022 pp. 113–136.

[28] Mullins, R., *Moral conflict and the logic of rights*, Philosophical Studies **177** (2020), pp. 633–51.

[29] Mullins, R., *Formalizing reasons, oughts, and requirements*, Ergo **7** (2021), pp. 568–99.

[30] Muñoz, D., *Three paradoxes of supererogation*, Noûs **55** (2021), pp. 669–716.

[31] Muñoz, D. and T. Pummer, *Supererogation and conditional obligation*, Philosophical Sstudies **179** (2022), pp. 1429–43.

[32] Nair, S., *Conflicting reasons, unconflicting 'oughts'*, Philosophical Studies **173** (2016), pp. 629–63.

[33] Parent, X. and L. van der Torre, "Introduction to Deontic Logic and Normative Systems," College Publications, 2018.

[34] Parfit, D., "On What Matters," Oxford University Press, 2011.

[35] Portmore, D., "Commonsense Consequentialism: Wherein Morality Meets Rationality," Oxford University Press, 2011.

[36] Portner, P., "Modality," Oxford University Press, 2009.

[37] Raz, J., "Practical reason and norms," Oxford University Press, 1990.

[38] Scanlon, T. M., "What We Owe to Each Other," Cambridge, MA: Harvard University Press, 1998.

[39] Scanlon, T. M., "Being Realistic about Reasons," Oxford University Press, 2014.

[40] Schroeder, M., "Slaves of the Passions," New York: Oxford University Press, 2007.

[41] Schroeder, M., "Reasons First," Oxford University Press, 2021.

[42] Silk, A., *What normative terms mean and why it matters for ethical theory*, Oxford Studies in Normative Ethics **25** (2015), pp. 296–325.

[43] Snedegar, J., *Reasons, oughts, and requirements*, Oxford Studies in Metaethcis **11** (2016), pp. 155–81.

[44] Star, D., *Introduction*, in: D. Star, editor, *The Oxford Handbook of Reasons and Normativity*, Oxford University Press, 2018 pp. 1–21.

[45] van Fraassen, B., *Values and the heart's command*, The Journal of Philosophy **70** (1973), pp. 5–19.

[46] von Fintel, K. and S. Iatridou, *Time and modality*, in: J. Guéron and J. Lecarme, editors, *How to say ought in foreign: The composition of weak necessity modals*, Springer-Verlag, 2008 pp. 115–41.

[47] Wessels, U., *Beyond the call of duty: The structure of a moral region*, Royal Institute of Philosophy Supplement: Supererogation **77** (2015), pp. 87–104.

Normative properties of sequential actions

Fengkui Ju [1]

School of Philosophy, Beijing Normal University
100875 Beijing, China

Karl Nygren [2]

Department of Philosophy, Stockholm University
SE-106 91 Stockholm, Sweden

Abstract

This paper develops a deontic logic based on dynamic logic for reasoning about permission and prohibition of sequential actions. Our approach is characterized by two main features. First, permission and prohibition of sequential actions are not necessarily reduced to permission and prohibition of the actions' constituent parts. Second, we incorporate the idea that actions may be permitted or prohibited conditional on another action being performed first. The logic interprets actions in terms of sequences of states, and the deontic component of the logic is introduced by relating sequences of states to their legal and illegal future continuations. We consider various different logics obtained by imposing natural constraints on models, and formulate complete axiom systems for several of these logics.

Keywords: Dynamic logic, sequential action, permission, prohibition

1 Introduction

This paper studies deontic logic for reasoning about permissions and prohibitions of sequential actions. Sequential actions are of the form "do α and then do β", and can be used to express procedures, plans and instructions which are executed by performing multiple actions in a step-by-step manner. Reasoning about normative properties of sequential actions is important for planning tasks in the presence of norms, where in addition to finding plans that achieve one's goals, one must also take into account which plans are permitted and which plans are prohibited.

Normative properties of sequential actions are usually studied using variants of *dynamic deontic logic*, i.e. variants of PDL extended with resources for reasoning about deontic concepts (see e.g. [2,5,6,9,10,11,13,14]). In the

[1] fengkui.ju@bnu.edu.cn

[2] karl.nygren@philosophy.su.se, corresponding author

dynamic deontic logic literature, there is a distinction between *goal-oriented* and *process-oriented* approaches [5]. Goal-oriented approaches define normative properties of actions in terms of their possible outcome states (e.g. [5,10]), whereas process-oriented approaches define normative properties of actions in terms of their possible executions (e.g. [6,9,11,14]). Here, we focus on the process-oriented approach. A central question in dynamic deontic logic is how "deontic properties of *compound* actions are logically related to deontic properties of *constituent parts* of actions" [6, p. 108]. When it comes to sequential actions, this question has typically been approached by defining normative properties of sequential actions in terms of normative properties of the transitions that occur during their possible executions (e.g. [1,7,11,14]), or in terms of the normative properties of atomic actions or "one-step" actions that occur when the sequential actions are performed (e.g. [6,9]).

In reality, however, normative properties of sequences of actions cannot always be reduced to combinations of normative properties of their constituent parts. For example, it is commonly forbidden to *drink and then drive* but not forbidden to *drive and then drink*. There seems to be no obvious way to obtain the normative properties of the two sequential actions by specifying normative properties of the action of drinking and the action of driving considered in isolation. In fact, this point has been explicitly argued by some literature outside deontic logic, e.g. [3,4]. Here we mention two interesting examples. The first one is from Bales and Benn [4, p. 7777]:

> Kwame is deciding how much weekly pocket money to give to his twin daughters. Any amount from nothing to £10 is an acceptable amount to give as pocket money. Given that any amount is acceptable, Kwame decides to give one daughter £1 and the other £10.

In this example, the sequential action of first giving one daughter £1 and then giving the other £10 is clearly problematic from a normative point of view. However, neither of the two actions in the sequence is problematic on its own. [3]

The second example is from Arntzenius, Elga and Hawthorne [3, p. 262]:

> Satan has cut a delicious apple into infinitely many pieces, labelled by the natural numbers. Eve may take whichever pieces she chooses. If she takes merely finitely many of the pieces, then she suffers no penalty. But if she takes infinitely many of the pieces, then she is expelled from the Garden for her greed.

In this example, each action of taking a piece of the apple is permissible, if considered individually. However, performing the sequential action of taking *every* piece is clearly impermissible: it leads to expulsion from the Garden.

Normative reasoning about sequential actions is closely connected to rea-

[3] Intuitively, the reason that the sequential action in the example is morally problematic is that one should not treat one's daughters unequally. Still, the example illustrates the problem with specifying the normative properties of a sequential action based on the normative properties of its constituent parts.

soning about what has already been done in the past. For example, from the fact that it is morally problematic for Kwame to give his second daughter £1 in a situation where he has already given his first daughter £10, we can naturally infer that the sequential action of giving the first daughter £1 followed by giving the second daughter £10 is morally problematic as well. Similarly, if you are permitted to finish your main course and then have dessert, it is natural to infer that if you have already finished your main course, you are permitted to have dessert. Taking a more 'future-looking' perspective, permissions and prohibitions that are dependent on past events can naturally be formulated in terms of permissions and prohibitions that hold *conditionally* on something else being done first. For example, a prohibition to drive a car after an intake of alcohol is naturally expressed in terms of a conditional prohibition: if you drink alcohol, you are prohibited from driving a car afterwards.

These considerations show that in order to properly account for permissions and prohibitions of sequential actions, we have to consider permissions and prohibitions that are dependent on past events. Alternatively, if we take a future-looking perspective, we have to consider permissions and prohibitions that hold conditionally on some other action being performed first.

In this paper, we develop a deontic logic based on dynamic logic for reasoning about permission and prohibition of sequential actions. The logic is characterized by two main features. First, the normative properties of sequential actions are not necessarily reduced to normative properties of their constituent parts. Second, the logic incorporates the idea that whether an action is permitted or prohibited may depend on past events. In the semantics for the logic, we interpret each action as a set of sequences of states, which intuitively correspond to the possible executions of the action. The deontic component of the logic is introduced in terms of relations that relate sequences of states (which we think of as incomplete records of past sequences of events) to their possible legal and illegal continuations. This allows us to specify the legal status of a whole sequence of states, without having to specify the legal status of the individual transitions occurring in it. In addition, by using relations between sequences of states, we allow for the legal status of a sequence of states to depend on which sequence of states it is a continuation of.

When interpreting the deontic operators, we take a future-looking perspective on the idea that the normative properties of actions may depend on past events. In particular, we consider dyadic deontic operators that specify an action to be permitted or prohibited conditional on another action being performed first. We consider a range of natural constraints that may be imposed on relations that relate past sequences of events to their possible legal or illegal continuations, and prove correspondences between these constraints and the validity of certain formulas. We also provide complete axiomatizations for several of the logics obtained by adopting various combinations of these constraints.

The paper makes a number of simplifying assumptions. For example, we only consider permission and prohibition, and leave the analysis of obligation of sequential actions to future research. Also, the language we use to talk about actions includes only the operations of sequential composition and non-deterministic choice, and is thus quite limited in its expressive powers.

The paper is structured as follows. The language and semantics of the logic are presented in Section 2.1 and Section 2.2, and an axiom system is provided in Section 2.3. In Section 3, we consider a range of natural constraints that can be imposed on models, and consider issues concerning correspondence results and axiomatizations. Section 4 concludes the paper. Some proofs can be found in the Appendix.

2 A deontic logic for reasoning about sequential actions

In this section, we define the language, semantics and axiomatic system for a dynamic deontic logic for reasoning about permission and prohibition of sequential actions.

2.1 Paths and path relations

Fix a non-empty set of states W. A *path* is a non-empty finite sequence of elements of W. Let W^* be the set of all paths consisting of elements of W. Given a path $\sigma = (w_1, \ldots, w_n)$, the first element w_1 of σ is denoted $\sigma[1]$ and the last element w_n of σ is denoted $\sigma[f]$. Given two paths $\sigma_1 = (w_1, \ldots, w_n)$ and $\sigma_2 = (v_1, \ldots, v_m)$, $\sigma_1 \circ \sigma_2$ denotes the path $(w_1, \ldots, w_n, v_2, \ldots, v_m)$ if $w_n = v_1$, otherwise $\sigma_1 \circ \sigma_2$ is undefined. We take paths to represent possible sequences of events, and we will later interpret actions as sets of paths.

A pair of paths (σ, σ') such that $\sigma \circ \sigma'$ is defined is called *an articulated history.*[4] Intuitively, we think of an articulated history (σ, σ') as representing a current state, a past, and a possible future in the following sense: $\sigma'[1] = \sigma[f]$ represents the current state, σ represents a past sequence of events, and σ' represents a possible future sequence of events.

A *path relation* defined on W is a set R of articulated histories, i.e. a set $\mathsf{R} \subseteq W^* \times W^*$ such that for each element $(\sigma, \sigma') \in \mathsf{R}$, $\sigma \circ \sigma'$ is defined.

2.2 Language and semantics

Let *Prop* be a countable set of atomic proposition symbols and *AtAct* be a countable set of atomic actions.

Definition 2.1 [Actions] The set *Act* of *actions* is defined by the following grammar, where a ranges over *AtAct*:

$$\alpha ::= a \mid \mathbf{skip} \mid \alpha \cup \alpha \mid \alpha ; \alpha .$$

An action of the form $\alpha \cup \beta$ expresses the *non-deterministic choice* between actions α and β. An action of the form $\alpha;\beta$ expresses the *sequential composition* of actions α and β, i.e. doing α followed by doing β. The action **skip** corresponds to the action of doing nothing.[5]

[4] We borrow the term 'articulated history' from Segerberg [13].

[5] We choose to work with a very simple action language here, and postpone the inclusion of other action operations, such as test actions, parallel execution and iteration, to future work.

Definition 2.2 [Language] The language $\mathcal{L}$ is defined by the following grammar, where p ranges over *Prop* and α and β range over *Act*:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid [\alpha]\varphi \mid \mathbf{P}(\alpha \mid \beta) \mid \mathbf{F}(\alpha \mid \beta).$$

The missing Boolean connectives are defined in the usual way. Formulas of the form $[\alpha]\varphi$ mean that each execution of α ends in state where φ is true. We use $\langle\alpha\rangle\varphi$ as an abbreviation of $\neg[\alpha]\neg\varphi$. Formulas of the form $\mathbf{P}(\alpha \mid \beta)$ and $\mathbf{F}(\alpha \mid \beta)$ express permissions and prohibitions that are *conditional* on another action being performed first: the intended interpretation of $\mathbf{P}(\alpha \mid \beta)$ is that after any execution of α, any way of doing β is legal, and the intended interpretation of $\mathbf{F}(\alpha \mid \beta)$ is that after some execution of α, there is some way of doing β which is illegal. Unary versions of $\mathbf{P}$ and $\mathbf{F}$ are defined by $\mathbf{P}\alpha := \mathbf{P}(\mathbf{skip} \mid \alpha)$ and $\mathbf{F}\alpha := \mathbf{F}(\mathbf{skip} \mid \alpha)$. $\mathbf{P}\alpha$ and $\mathbf{F}\alpha$ mean that α is permitted, respectively prohibited, *immediately*: nothing else has to be done first.

We interpret each action α as a set of paths. When a path σ belongs to the interpretation of an action α, it intuitively means that σ constitutes a possible execution of α.

Definition 2.3 [Interpretation of actions] Given a set of states W, an *action interpretation* is a function I assigning a set of paths $I(a) \subseteq W^*$ to each $a \in AtAct$. I is extended to cover all actions in *Act* in the following way:

$$\begin{aligned} I(\mathbf{skip}) &:= W; \\ I(\alpha \cup \beta) &:= I(\alpha) \cup I(\beta); \\ I(\alpha;\beta) &:= \{\sigma \circ \sigma' \mid \sigma \in I(\alpha), \sigma' \in I(\beta), \sigma'[1] = \sigma[f]\}. \end{aligned}$$

To interpret permission and prohibition, we introduce path relations LEG and ILL that encode *legal* respectively *illegal* paths: $(\sigma, \sigma') \in \mathsf{LEG}$ means that the path σ' is legal given that it occurs after σ has been realized, and $(\sigma, \sigma') \in \mathsf{ILL}$ means that the path σ' is illegal given that it occurs after σ has been realized. If $(\sigma, \sigma') \in \mathsf{LEG}/\mathsf{ILL}$, we say that σ' is a legal, respectively illegal, *continuation* of σ.

One may think of LEG and ILL as being determined by a normative system. We assume that the function of a normative system is to determine, given some past sequence of events, which possible future sequences of events are legal and which possible future sequences of events are illegal; cf. [1,7,13]. Under this assumption, the sets $\mathsf{LEG}(\sigma) = \{\sigma' \mid (\sigma, \sigma') \in \mathsf{LEG}\}$ and $\mathsf{ILL}(\sigma) = \{\sigma' \mid (\sigma, \sigma') \in \mathsf{ILL}\}$ consist of those possible continuations of σ that are considered legal, respectively illegal, by the normative system.

To keep things as general as possible, we do not require any interaction principles between LEG and ILL. Consequently, our models allow for paths that are neither legal nor illegal, and paths that are simultaneously both legal and illegal. This opens up possibilities for modeling normative systems with gaps, as well as contradictory normative systems; cf. [16, p. 32]. In Section 3, we consider additional constraints that may be imposed to rule out gappy and contradictory normative systems.

By using path relations, we do not have to assume that the legal status of a path is completely determined by the legal status of its subpaths. Rather, our approach makes room for several different options: on one extreme, we can model cases where the legal status of a path is completely independent of the legal status of its subpaths; on the other extreme, we can model cases where the legal status of a path is completely determined by the legal status of its subpaths. We can also model positions in between these two extremes, where certain normative properties of the parts of a path may constitute necessary or sufficient conditions for the path to be legal or illegal.

The use of path relations also allows for the legal status of a path to be dependent on a past sequence of events. In principle, this means that a path may be considered legal in relation to one past sequence of events, and illegal in relation to another.

Definition 2.4 [Frames and models] A *frame* is a structure $\mathcal{F} = (W, \mathsf{LEG}, \mathsf{ILL})$, where W is a non-empty set of states and LEG and ILL are path relations defined on W. A *model* $\mathcal{M} = (W, \mathsf{LEG}, \mathsf{ILL}, I, V)$ based on a frame $\mathcal{F} = (W, \mathsf{LEG}, \mathsf{ILL})$ extends $\mathcal{F}$ with an action interpretation I assigning a set of paths $I(a) \subseteq W^*$ to each $a \in \mathit{AtAct}$, and a function V assigning a set of states $V(p) \subseteq W$ to each $p \in \mathit{Prop}$.

Definition 2.5 [Interpretation of formulas] Formulas are interpreted relative to a model $\mathcal{M} = (W, \mathsf{LEG}, \mathsf{ILL}, I, V)$ and a state w according to the following clauses: [6]

$$\begin{array}{lll}
\mathcal{M}, w \models p & \text{iff} & w \in V(p); \\
\mathcal{M}, w \models \neg\varphi & \text{iff} & \mathcal{M}, w \not\models \varphi; \\
\mathcal{M}, w \models \varphi \wedge \psi & \text{iff} & \mathcal{M}, w \models \varphi \text{ and } \mathcal{M}, w \models \psi; \\
\mathcal{M}, w \models [\alpha]\varphi & \text{iff} & \text{for all } \sigma \in I(\alpha), \text{ if } \sigma[1] = w \text{ then } \mathcal{M}, \sigma[f] \models \varphi; \\
\mathcal{M}, w \models \mathbf{P}(\alpha \mid \beta) & \text{iff} & \text{for all } \sigma \in I(\alpha), \text{ for all } \sigma' \in I(\beta), \\
 & & \text{if } \sigma[1] = w \text{ and } \sigma'[1] = \sigma[f], \text{ then } (\sigma, \sigma') \in \mathsf{LEG}; \\
\mathcal{M}, w \models \mathbf{F}(\alpha \mid \beta) & \text{iff} & \text{there is } \sigma \in I(\alpha) \text{ and there is } \sigma' \in I(\beta) \\
 & & \text{such that } \sigma[1] = w, \sigma'[1] = \sigma[f] \text{ and } (\sigma, \sigma') \in \mathsf{ILL}.
\end{array}$$

Definition 2.6 [Validity and logical consequence] A formula $\varphi \in \mathcal{L}$ is *valid*, notation $\models \varphi$, if $\mathcal{M}, w \models \varphi$ for all models $\mathcal{M}$ and all states w of $\mathcal{M}$, and φ is *valid in a frame* $\mathcal{F}$, notation $\mathcal{F} \models \varphi$, if $\mathcal{M}, w \models \varphi$ for all models $\mathcal{M}$ based on $\mathcal{F}$ and all states w of $\mathcal{M}$. A formula $\psi \in \mathcal{L}$ is a *logical consequence* of a set of formulas $\Phi \subseteq \mathcal{L}$ if for all models $\mathcal{M}$ and all states w of $\mathcal{M}$, if $\mathcal{M}, w \models \varphi$ for all $\varphi \in \Phi$, then $\mathcal{M}, w \models \psi$.

One can think of a formula of the form $\mathbf{P}(\alpha \mid \beta)$ as expressing a permission to freely choose any of the possible ways to do β after having performed α. This means that the permission operator $\mathbf{P}$ captures a notion of 'strong' or

[6] Recall that the action interpretation function I is extended to cover all actions in *Act* as in Definition 2.3.

'free choice' permission [8,15]. The following validities hold:

$$\mathbf{P}(\delta \mid \alpha \cup \beta) \leftrightarrow \mathbf{P}(\delta \mid \alpha) \wedge \mathbf{P}(\delta \mid \beta);$$
$$\mathbf{P}(\alpha \cup \beta \mid \delta) \leftrightarrow \mathbf{P}(\alpha \mid \delta) \wedge \mathbf{P}(\beta \mid \delta).$$

A formula of the form $\mathbf{F}(\alpha \mid \beta)$ can be understood as expressing that freely choosing how to perform β after having done α is prohibited: if there is some illegal way of doing β after α is performed, then any choice that includes that way of doing β as an option is prohibited (cf. [6, p. 108] and [5, p. 165]). This way of interpreting prohibition results in the following validities:

$$\mathbf{F}(\delta \mid \alpha \cup \beta) \leftrightarrow \mathbf{F}(\delta \mid \alpha) \vee \mathbf{F}(\delta \mid \beta);$$
$$\mathbf{F}(\alpha \cup \beta \mid \delta) \leftrightarrow \mathbf{F}(\alpha \mid \delta) \vee \mathbf{F}(\beta \mid \delta).$$

The first principle says that it is prohibited to choose between α and β conditional on doing δ first if and only if it is prohibited to do α after doing δ or it is prohibited to do β after doing δ. The second property says that being prohibited to do δ after making a choice between α and β is equivalent to prohibition to do δ after doing α or prohibition to do δ after doing β.

2.3 Axiomatization

We start by considering a language for *action equivalence*, consisting of expressions of the form $\alpha \equiv \beta$, where α and β are actions from Act. The formal semantics of these types of expressions are given as follows: $\alpha \equiv \beta$ is true in a model $\mathcal{M}$ if and only if $I(\alpha) = I(\beta)$, where I is the action interpretation function of $\mathcal{M}$. We say that $\alpha \equiv \beta$ is *valid* if it is true in all models. An axiomatization of action equivalence is given by the following axioms and inference rules:

- Axioms:
  (A1) $\alpha \equiv \alpha$
  (A2) $\alpha \cup \alpha \equiv \alpha$
  (A3) $\alpha \cup \beta \equiv \beta \cup \alpha$
  (A4) $\alpha \cup (\beta \cup \delta) \equiv (\alpha \cup \beta) \cup \delta$ and $\alpha; (\beta; \delta) \equiv (\alpha; \beta); \delta$
  (A5) $\alpha; (\beta \cup \delta) \equiv (\alpha; \beta) \cup (\alpha; \delta)$ and $(\alpha \cup \beta); \delta \equiv (\alpha; \delta) \cup (\beta; \delta)$
  (A6) $\mathbf{skip}; \alpha \equiv \alpha$ and $\alpha; \mathbf{skip} \equiv \alpha$
- From $\alpha \equiv \beta$, infer $\beta \equiv \alpha$.
- From $\alpha \equiv \beta$, infer $\delta \equiv \delta'$, where δ' is obtained from δ by replacing some or all occurrences of α in δ by β.

An action is in *normal form* if it is of the form $\delta_1 \cup \cdots \cup \delta_n$, where for each δ_i, either $\delta_i = \mathbf{skip}$, or δ_i is a sequence of atomic actions with association to the right, and for each δ_i and δ_j, if $i \neq j$ then $\delta_i \neq \delta_j$. It can be shown that any action is both semantically and provably equivalent to one in normal form. Using this fact, we can prove the following soundness and completeness result:

Theorem 2.7 *For any $\alpha, \beta \in Act$, $\alpha \equiv \beta$ is derivable in the above axiomatic system iff $\alpha \equiv \beta$ is valid.*

Next, we use the axiomatization of action equivalence to formulate an axiom system for our deontic logic, given by the following axioms and inference rules:

- Axioms for propositional logic.
- Axioms for modal operators:
  (M1) $[a](\varphi \to \psi) \to ([a]\varphi \to [a]\psi)$
  (M2) $[a \cup b]\varphi \leftrightarrow ([a]\varphi \wedge [b]\varphi)$
  (M3) $[a; b]\varphi \leftrightarrow [a][b]\varphi$
  (M4) $[\mathbf{skip}]\varphi \leftrightarrow \varphi$
- Axioms for deontic operators:
  (D1) $\mathbf{P}(c \mid a \cup b) \leftrightarrow (\mathbf{P}(c \mid a) \wedge \mathbf{P}(c \mid b))$
  (D2) $\mathbf{P}(a \cup b \mid c) \leftrightarrow (\mathbf{P}(a \mid c) \wedge \mathbf{P}(b \mid c))$
  (D3) $\neg\mathbf{P}(a \mid b) \to \langle a; b\rangle\top$
  (D4) $\mathbf{F}(c \mid a \cup b) \leftrightarrow (\mathbf{F}(c \mid a) \vee \mathbf{F}(c \mid b))$
  (D5) $\mathbf{F}(a \cup b \mid c) \leftrightarrow (\mathbf{F}(a \mid c) \vee \mathbf{F}(b \mid c))$
  (D6) $\mathbf{F}(a \mid b) \to \langle a; b\rangle\top$
- Modus ponens: from φ and $\varphi \to \psi$, infer ψ.
- Action replacement: from $\alpha \equiv \beta$, infer $D(\alpha \mid \delta) \leftrightarrow D(\beta \mid \delta)$ and $D(\delta \mid \alpha) \leftrightarrow D(\delta \mid \beta)$, where $D \in \{\mathbf{P}, \mathbf{F}\}$.
- Necessitation: from φ, infer $[a]\varphi$.
- Uniform action substitution: from φ, infer θ, where θ is obtained from φ by uniformly replacing atomic actions in φ by arbitrary actions.[7]

We use $\vdash \psi$ to mean that ψ is derivable in the above axiomatic system. For any $\Phi \subseteq \mathcal{L}$ and any $\psi \in \mathcal{L}$, $\Phi \vdash \psi$ means that $\vdash \varphi_1 \wedge \cdots \wedge \varphi_n \to \psi$ for some $\varphi_1, \ldots, \varphi_n \in \Phi$.

Proving soundness is routine. We prove completeness in Appendix A.

Theorem 2.8 (Soundness and completeness) *For any $\Phi \subseteq \mathcal{L}$ and any $\psi \in \mathcal{L}$, $\Phi \models \psi$ iff $\Phi \vdash \psi$.*

3 Additional constraints on path relations

In this section, we consider some natural constraints that may be imposed on the path relations LEG and ILL, and discuss the logical principles that correspond to these constraints. The different constraints can be freely combined to form various different deontic logics to pick and choose from, depending on one's modeling needs. We show correspondences between the constraints on LEG and ILL and the validity of certain formulas, and provide complete axiomatizations for several of the logics obtained by adopting various combinations of these constraints.

[7] We note that the rule of uniform action substitution is non-standard in dynamic logic; see [12] for some discussion.

3.1 Constraints on the interaction between LEG and ILL

A natural constraint on the interaction between LEG and ILL is to require that no path is both legal and illegal. This constraint is suitable when modeling normative systems where no conflicts between norms arise [16, p. 32]. Accordingly, we call it the *no conflicts* constraint:

$$\textbf{NoCon} \qquad \mathsf{LEG} \cap \mathsf{ILL} = \emptyset.$$

Adopting **NoCon** results in the following validity: [8]

$$\mathbf{P}(\alpha \mid \beta) \to \neg\mathbf{F}(\alpha \mid \beta).$$

A second natural constraint is to require that there are no gaps in the division of paths into legal and illegal ones: everything which is not illegal is legal [16, p. 32]. In our framework, this is captured by the following property (we assume implicit universal quantification over paths):

$$\textbf{NoGap} \qquad (\sigma, \sigma') \notin \mathsf{ILL} \text{ and } \sigma'[1] = \sigma[f] \text{ implies } (\sigma, \sigma') \in \mathsf{LEG}.$$

Imposing **NoGap** on models results in the following validity:

$$\neg\mathbf{F}(\alpha \mid \beta) \to \mathbf{P}(\alpha \mid \beta).$$

3.2 Constraints on LEG

We say that a path relation R is *closed under forward motion* if the following condition holds:

$$\textbf{FoMo} \qquad (\sigma, \sigma' \circ \sigma'') \in \mathsf{R} \text{ implies } (\sigma \circ \sigma', \sigma'') \in \mathsf{R}.$$

Segerberg [13, p. 393] refers to a similar property as the *coherence condition*: when imposed on the relation LEG, **FoMo** can be understood as capturing the idea that a legal path is 'coherent' in the sense that its legal status does not change in the middle of it. At the level of validities, **FoMo** captures the property that if a sequential action $\alpha;\beta$ is permitted conditional on performing δ, then β is permitted conditional on performing $\delta;\alpha$. Closing LEG under **FoMo** corresponds to validating the following principle:

$$\mathbf{P}(\delta \mid \alpha;\beta) \to \mathbf{P}(\delta;\alpha \mid \beta).$$

A path relation R is *closed under future initial segments* if the following condition holds:

$$\textbf{InSeg} \qquad (\sigma, \sigma' \circ \sigma'') \in \mathsf{R} \text{ implies } (\sigma, \sigma') \in \mathsf{R}.$$

When imposed on the relation LEG, we may think of **InSeg** as capturing the property that if a sequential action is permitted, then the first part of the

[8] All constraints that we consider in this and the following subsections, together with their corresponding validities, are listed in Theorem 3.1 below.

sequential action is also permitted. Closing LEG under **InSeg** corresponds to validating the following principle:

$$\mathbf{P}(\delta \mid \alpha;\beta) \wedge [\delta;\alpha]\langle\beta\rangle\top \to \mathbf{P}(\delta \mid \alpha).$$

The additional conjunct $[\delta;\alpha]\langle\beta\rangle\top$ in the antecedent is needed to rule out cases where there are executions of α ending in states where β cannot be performed.

A path relation R is *closed under path-transitivity* if the following condition holds:

$$\textbf{Trans} \qquad (\sigma,\sigma') \in \mathsf{R} \text{ and } (\sigma\circ\sigma',\sigma'') \in \mathsf{R} \text{ implies } (\sigma,\sigma'\circ\sigma'') \in \mathsf{R}.$$

When imposed on the relation LEG, **Trans** captures the idea that if an action α is permitted conditional on performing δ, and the action β is permitted conditional on performing $\delta;\alpha$, then the action $\alpha;\beta$ is permitted conditional on performing δ. This is mirrored by the fact that closing LEG under **Trans** corresponds to validating the following principle:

$$\mathbf{P}(\delta \mid \alpha) \wedge \mathbf{P}(\delta;\alpha \mid \beta) \to \mathbf{P}(\delta \mid \alpha;\beta).$$

The principle lays down sufficient conditions for inferring the permission of a sequential action from the permission of its parts.

We note in particular that if LEG is closed under **FoMo**, **InSeg** and **Trans** at the same time, then the following principle holds:

$$[\delta;\alpha]\langle\beta\rangle\top \to (\mathbf{P}(\delta \mid \alpha;\beta) \leftrightarrow \mathbf{P}(\delta \mid \alpha) \wedge \mathbf{P}(\delta;\alpha \mid \beta)).$$

This principle can be seen as a kind of *reduction principle* for permission: under some conditions (given by the antecedent of the above schema), being permitted to perform a sequential action is equivalent to being permitted to perform each part of the sequential action.

3.3 Constraints on ILL

We now turn to properties that we think are natural to impose on the relation ILL. The first property is called *closure under backward motion*; a path relation R satisfies this property if the following condition holds:

$$\textbf{BaMo} \qquad (\sigma\circ\sigma',\sigma'') \in \mathsf{R} \text{ implies } (\sigma,\sigma'\circ\sigma'') \in \mathsf{R}.$$

Closing the relation ILL under **BaMo** corresponds the following validity:

$$\mathbf{F}(\delta;\alpha \mid \beta) \to \mathbf{F}(\delta \mid \alpha;\beta).$$

That is, if an action β is prohibited after the action $\delta;\alpha$ has been performed, then the sequential action $\alpha;\beta$ is prohibited after δ has been performed. This seems like an intuitively desirable property. From the prohibition to drive a car after drinking alcohol, it intuitively follows that the sequential action of drinking alcohol followed by driving a car is prohibited as well.

Next, a path relation R is *closed under forward extension* if the following condition holds:

$$\textbf{FoEx} \qquad (\sigma, \sigma') \in \mathsf{R} \text{ implies } (\sigma, \sigma' \circ \sigma'') \in \mathsf{R}.$$

Imposing **FoEx** on the relation ILL captures the property that if some path σ is illegal, then any path $\sigma \circ \sigma'$ extending σ is also illegal. If ILL is closed under **FoEx**, then the following *principle of excluded Robin Hood* is valid: [9]

$$\mathbf{F}(\delta \mid \alpha) \wedge [\delta; \alpha]\langle \beta \rangle \top \rightarrow \mathbf{F}(\delta \mid \alpha; \beta).$$

As before, the additional conjunct $[\delta; \alpha]\langle \beta \rangle \top$ in the antecedent is needed to rule out cases where there are executions of α ending in states where β cannot be performed. This principle says that if some action α is prohibited (conditional on performing δ first), then performing α followed by some other action β is also prohibited (conditional on performing δ first). This can be understood as capturing the idea that an action can never be made "better" by performing some other action afterwards.

Next, we consider the following property: a path relation R is *closed under splitting paths* if the following condition hold:

$$\textbf{Split} \qquad (\sigma, \sigma' \circ \sigma'') \in \mathsf{R} \text{ implies } (\sigma, \sigma') \in \mathsf{R} \text{ or } (\sigma \circ \sigma', \sigma'') \in \mathsf{R}.$$

When imposed on the relation ILL, this property corresponds to the idea that a necessary condition for a path to be illegal is that there is some illegal transition occurring in it. When closing ILL under **Split**, the following principle is valid:

$$\mathbf{F}(\delta \mid \alpha; \beta) \rightarrow \mathbf{F}(\delta \mid \alpha) \vee \mathbf{F}(\delta; \alpha \mid \beta).$$

That is, if $\alpha; \beta$ is prohibited after some execution of δ, it follows that either α is prohibited after some execution of δ, or β is prohibited after some execution of $\delta; \alpha$.

If ILL is closed under all three of **BaMo**, **FoEx** and **Split**, the following principle holds:

$$[\delta; \alpha]\langle \beta \rangle \top \rightarrow (\mathbf{F}(\delta \mid \alpha; \beta) \leftrightarrow \mathbf{F}(\delta \mid \alpha) \vee \mathbf{F}(\delta; \alpha \mid \beta)).$$

This principle can be seen as a reduction principle for prohibition: under the conditions specified in the antecedent, the principle tells us how a prohibited sequential action can be reduced to a combination of prohibitions of its constituent parts.

[9] The name comes from the following instance: "If it is forbidden to take money from the rich, then it is forbidden to take money from the rich and give it to the poor afterwards." [2, p. 429].

3.4 Properties concerning the past

We consider two additional properties, which may be imposed on LEG and ILL: *closure under past final segments* and *closure under backwards extension*:

$$\begin{array}{ll} \textbf{FiSeg} & (\sigma \circ \sigma', \sigma'') \in \mathsf{R} \text{ implies } (\sigma', \sigma'') \in \mathsf{R}; \\ \textbf{BaEx} & (\sigma', \sigma'') \in \mathsf{R} \text{ implies } (\sigma \circ \sigma', \sigma'') \in \mathsf{R}. \end{array}$$

The property **FiSeg** can be understood as cutting off a part of the past of an articulated history, whereas **BaEx** can be understood as extending the past of an articulated history backwards.

When imposed on the relations LEG and ILL, **FiSeg** captures the property that if a path σ' is a legal/illegal continuation of a path σ, then σ' is a legal/illegal continuation of any final segment of σ as well. This property corresponds to the idea that if the action β is permitted/prohibited after $\delta; \alpha$ has been performed, then β is permitted/prohibited after α has been performed. If LEG, respectively ILL, are closed under **FiSeg**, then the following validities hold:

$$\begin{gathered} \mathbf{P}(\delta; \alpha \mid \beta) \rightarrow [\delta]\mathbf{P}(\alpha \mid \beta) \\ \mathbf{F}(\delta; \alpha \mid \beta) \rightarrow \langle\delta\rangle\mathbf{F}(\alpha \mid \beta). \end{gathered}$$

The constraint **BaEx**, when imposed on LEG or ILL, captures the property that if a path σ' is a legal/illegal continuation of σ, then σ' is a legal/illegal continuation of any path σ'' of which σ is a final segment. This property can be taken to mean that if the action β is permitted/prohibited conditional on α being performed first, then β is permitted/prohibited conditional on α being performed first, regardless of what happens before α is performed. If LEG, respectively ILL, are closed under **BaEx**, then the following validities hold:

$$\begin{gathered} [\delta]\mathbf{P}(\alpha \mid \beta) \rightarrow \mathbf{P}(\delta; \alpha \mid \beta) \\ \langle\delta\rangle\mathbf{F}(\alpha \mid \beta) \rightarrow \mathbf{F}(\delta; \alpha \mid \beta). \end{gathered}$$

We note in particular that if LEG and ILL satisfy both **FiSeg** and **BaEx**, the dyadic deontic operators can be reduced to a combination of unary deontic operators and dynamic modal operators in the following way (recall that $\mathbf{P}\alpha$ abbreviates $\mathbf{P}(\mathbf{skip} \mid \alpha)$ and that $\mathbf{F}\alpha$ abbreviates $\mathbf{F}(\mathbf{skip} \mid \alpha)$):

$$\begin{gathered} \mathbf{P}(\alpha \mid \beta) \leftrightarrow [\alpha]\mathbf{P}\beta \\ \mathbf{F}(\alpha \mid \beta) \leftrightarrow \langle\alpha\rangle\mathbf{F}\beta. \end{gathered}$$

3.5 Corresponding validities and axiomatization

The following theorem connects the different path relation properties with their corresponding validities. We prove a representative example in Appendix B.

Theorem 3.1 *Let $\mathcal{F} = (W, \mathsf{LEG}, \mathsf{ILL})$ be a frame. Then the following hold:*

(i) LEG *and* ILL *together satisfy* **NoCon** *iff* $\mathcal{F} \models \mathbf{P}(a \mid b) \rightarrow \neg\mathbf{F}(a \mid b)$*;*

(ii) LEG *and* ILL *together satisfy* **NoGap** *iff* $\mathcal{F} \models \neg\mathbf{F}(a \mid b) \to \mathbf{P}(a \mid b)$;

(iii) LEG *is closed under* **FoMo** *iff* $\mathcal{F} \models \mathbf{P}(a \mid b; c) \to \mathbf{P}(a; b \mid c)$;

(iv) LEG *is closed under* **InSeg** *iff* $\mathcal{F} \models \mathbf{P}(a \mid b; c) \wedge [a; b]\langle c\rangle\top \to \mathbf{P}(a \mid b)$;

(v) LEG *is closed under* **Trans** *iff* $\mathcal{F} \models \mathbf{P}(a \mid b) \wedge \mathbf{P}(a; b \mid c) \to \mathbf{P}(a \mid b; c)$;

(vi) LEG *is closed under* **FiSeg** *iff* $\mathcal{F} \models \mathbf{P}(a; b \mid c) \to [a]\mathbf{P}(b \mid c)$;

(vii) LEG *is closed under* **BaEx** *iff* $\mathcal{F} \models [a]\mathbf{P}(b \mid c) \to \mathbf{P}(a; b \mid c)$;

(viii) ILL *is closed under* **BaMo** *iff* $\mathcal{F} \models \mathbf{F}(a; b \mid c) \to \mathbf{F}(a \mid b; c)$;

(ix) ILL *is closed under* **FoEx** *iff* $\mathcal{F} \models \mathbf{F}(a \mid b) \wedge [a; b]\langle c\rangle\top \to \mathbf{F}(a \mid b; c)$;

(x) ILL *is closed under* **Split** *iff* $\mathcal{F} \models \mathbf{F}(a \mid b; c) \to \mathbf{F}(a \mid b) \vee \mathbf{F}(a; b \mid c)$;

(xi) ILL *is closed under* **FiSeg** *iff* $\mathcal{F} \models \mathbf{F}(a; b \mid c) \to \langle a\rangle\mathbf{F}(b \mid c)$;

(xii) ILL *is closed under* **BaEx** *iff* $\mathcal{F} \models \langle a\rangle\mathbf{F}(b \mid c) \to \mathbf{F}(a; b \mid c)$.

What about the logics concerning these properties? The properties **NoCon** and **NoGap** for LEG and ILL, the properties **FoMo**, **Trans** and **FiSeg** for LEG, and the properties **BaMo**, **Split** and **BaEx** for ILL can be directly handled by the approach to proving completeness of the logic given in Section 2.3. To be precise, we have the following claim (the proof of the soundness part of the theorem is routine; we sketch the proof of the completeness part of the theorem in Appendix C):

Theorem 3.2 *For every property* Pr *of* LEG *and* ILL *in* $\{\mathbf{NoCon}, \mathbf{NoGap}\}$*, we use* ϕ^{PF}_{Pr} *to indicate the corresponding formula, for every property* Pr *of* LEG *in* $\{\mathbf{FoMo}, \mathbf{Trans}, \mathbf{FiSeg}\}$*, we use* ϕ^{P}_{Pr} *to indicate the corresponding formula, and for every property* Pr *of* ILL *in* $\{\mathbf{BaMo}, \mathbf{Split}, \mathbf{BaEx}\}$*, we use* ϕ^{F}_{Pr} *to indicate the corresponding formula. Let* Φ *be the set consisting of the eight formulas. For every nonempty subset* Φ' *of* Φ*, the logic obtained by adding the formulas in* Φ' *as axioms to the logic given in Section 2.3 is sound and complete with respect to the class of models with the corresponding properties.*

The other four properties cannot be dealt with directly.

4 Concluding remarks

In this paper, we introduced and studied a deontic logic for reasoning about permission and prohibition of sequential actions. In the semantics for the logic, we interpret actions as sets of paths, which intuitively correspond to the possible executions of the actions. The deontic component of the logic is introduced in terms of relations between paths, which intuitively relate paths representing possible past sequences of events to their possible legal and illegal continuations. The logic features dyadic deontic operators that can be used to express that an action is permitted, respectively prohibited, conditional on some other action being performed first. We provided the logic with a complete axiomatization, and studied several natural constraints that may be imposed on relations between paths in order to validate additional deontic logic principles.

There are several important topics for future work. In addition to permissions and prohibitions, we would like to consider obligations of sequential actions. One complication is that obligation intuitively involves a notion of omission (e.g. α is obligatory if and only if it is prohibited to omit doing α), and formalizing what it means to omit doing a sequential action is notoriously difficult [5]. One option that seems promising is to introduce obligation as a primitive notion; for work in this direction, see [9].

In this work, we interpret permission and prohibition using universal-universal and existential-existential quantifier patterns, respectively (that is, $\mathbf{P}(\alpha \mid \beta)$ is true iff after *all* α-paths, *all* β-paths are legal, and $\mathbf{F}(\alpha \mid \beta)$ is true iff after *some* α-path, *some* β-path is legal). It would also be natural to interpret permission using an existential-existential quantifier pattern (thus obtaining a kind of 'weak' permission concept [15]), and to interpret prohibition using a universal-universal quantifier pattern (which would be more in line with established tradition in deontic logic). In addition, there are two other quantifier patterns to consider for the interpretation of permission and prohibition: universal-existential and existential-universal. An interesting direction for future research is to study the concepts of permission and prohibition arising from these different quantifier patterns. One option is to consider all of these different concepts of permission and prohibition in the same logic, thus obtaining a more expressive language for reasoning about normative properties of sequential actions.

We would also like to investigate more closely the idea that the relations LEG and ILL can be understood as being determined by a normative system. In [7], a formal model of norms and normative systems is developed and used to determine the legal status of transitions in a transition system. However, only norms that regulate non-sequential actions were considered in that paper, and it seems that the formal model must be extended in order to handle norms that regulate sequential actions directly.

In this work, we take a 'future-looking' perspective on the idea that the normative properties of actions may depend on the past. This can be seen from the readings of the two deontic operators: after any/some execution of α, any/some way of doing β is permitted/prohibited. There is also a 'past-future-looking' perspective on the idea, that is, to introduce the following deontic operators: given that α has been executed, any/some way of doing β is permitted/prohibited. This is similar to the approach in Segerberg's dynamic deontic logic [13], where deontic formulas are evaluated at paths representing the past, rather than single states. We leave the investigation of this approach, as well as a detailed comparison with Segerberg's logic, for future work.

On the technical side, we would like to establish complete axiomatizations for all logics obtained by imposing any of the constraints on path relations considered in Section 3. As shown in Section 3.5, for some of these logics we can obtain complete axiomatizations by adopting the formulas corresponding to the path relation properties as axioms. However, this strategy does not generalize directly to all path relation properties, and further work is needed.

Acknowledgement

We thank Valentin Goranko, Daniel Ramöller, the anonymous reviewers, and the audience of a workshop at Chinese Academy of Sciences for helpful discussions, comments and suggestions. Fengkui Ju was supported by the National Social Science Foundation of China (No. 19BZX137).

Appendix

A Completeness proof

Here, we sketch the proof of the completeness part of Theorem 2.8. First, we provide some definitions and establish some lemmata.

Definition A.1 The language $\mathcal{L}_{red}$ is defined by the following grammar, where p ranges over *Prop* and a ranges over *AtAct*:

$$\begin{aligned}\varphi &::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid [a]\varphi \mid \mathbf{P}(\delta \mid \delta) \mid \mathbf{F}(\delta \mid \delta)\\ \delta &::= \mathbf{skip} \mid \gamma\\ \gamma &::= a \mid \gamma;\gamma\end{aligned}$$

In the language $\mathcal{L}_{red}$, only atomic actions occur in dynamic modalities, and any action within the scope of a deontic operator is either in the form of a sequence of atomic actions, or equal to **skip**.

Lemma A.2 *There is a translation $t : \mathcal{L} \to \mathcal{L}_{red}$ such that for each formula $\varphi \in \mathcal{L}$, $\vdash \varphi \leftrightarrow t(\varphi)$.*

Let Ω be the set of maximally consistent sets of formulas of $\mathcal{L}_{red}$. For each $a \in AtAct$, we define the relation $R_a \subseteq \Omega \times \Omega$ as follows:

$$(\Delta, \Delta') \in R_a \quad \text{iff} \quad \text{for all } \varphi \in \mathcal{L}_{red}, \text{ if } [a]\varphi \in \Delta \text{ then } \varphi \in \Delta'.$$

Fix a maximally consistent set Δ_0. We define the set W^c step by step:

- **Step** 0. Let $w_0 = (0, \Delta_0)$ and set $W^c_0 = \{w_0\}$.
- **Step** 1. For each maximally consistent set $\Delta \in \Omega$ and each $a \in AtAct$ such that $(\Delta_0, \Delta) \in R_a$, put the point $v = (1, w_0, a, \Delta)$ in W^c_1.
- **Step** $i+1$. Assume that the set W^c_i has been defined. For each maximally consistent set $\Delta \in \Omega$, each action $a \in AtAct$ and each $w = (i, w', a', \Delta') \in W^c_i$ such that $(\Delta', \Delta) \in R_a$, put the point $v = (i+1, w, a, \Delta)$ in W^c_{i+1}.

Finally, we set W^c to be the union of all W^c_i for $i \in \mathbb{N}$.

By construction of W^c, each element w is associated with a maximally consistent set which we denote $\mathsf{mcs}(w)$: for $w = (0, \Delta_0) \in W^c_0$, set $\mathsf{mcs}(w) = \Delta_0$; for $w = (i, v, a, \Delta) \in W^c_i$, with $i \geq 1$, set $\mathsf{mcs}(w) = \Delta$. For each element $w = (i, v, a, \Delta) \in W^c_i$ with $i \geq 1$, we say that v is the *predecessor* of w and denote it by $\mathsf{pre}(w)$, and that a is the *action label* of w and denote it by $\mathsf{act}(w)$.

Define the interpretation $I^c(a)$ of the atomic action $a \in AtAct$ as a set of paths of length 2 in the following way:

$$I^c(a) = \{(w, w') \in W^c \times W^c \mid \mathsf{act}(w') = a \text{ and } \mathsf{pre}(w') = w\}.$$

For any path $\sigma \in (W^c)^*$ such that w_0 either only occurs as the first element of σ, or w_0 does not occur in σ at all, we define the *action trace* of σ recursively as follows:

- $\mathsf{trace}(w) = \mathbf{skip}$
- $\mathsf{trace}((w, w')) = \mathsf{act}(w')$
- $\mathsf{trace}((w, w') \circ \tau) = \mathsf{act}(w'); \mathsf{trace}(\tau)$, if τ is of length ≥ 2.

Lemma A.3 *Let δ be either equal to* **skip**, *or equal to a sequence of atomic actions with association to the right. If $\sigma \in I^c(\delta)$ then $\mathsf{trace}(\sigma) = \delta$.*

Lemma A.4 *Let $\gamma = a_1; \ldots; a_k$ be a sequence of atomic actions with association to the right, let $w \in W^c$ and let $\varphi \in \mathcal{L}_{red}$. If $\mathsf{mcs}(w) \vdash \langle a_1 \rangle \ldots \langle a_k \rangle \varphi$, then there is $\sigma \in I^c(\gamma)$ such that $\sigma[1] = w$ and $\mathsf{mcs}(\sigma[f]) \vdash \varphi$.*

Next, we define the path relations LEG^c and ILL^c on W^c as follows:

$$\begin{array}{lll} (\sigma, \sigma') \in \mathsf{LEG}^c & \text{iff} & \sigma'[1] = \sigma[f] \text{ and } \mathsf{mcs}(\sigma[1]) \vdash \mathbf{P}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma')); \\ (\sigma, \sigma') \in \mathsf{ILL}^c & \text{iff} & \sigma'[1] = \sigma[f] \text{ and } \mathsf{mcs}(\sigma[1]) \vdash \mathbf{F}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma')). \end{array}$$

We define the interpretation V^c for atomic proposition symbols by setting $w \in V^c(p)$ iff $p \in \mathsf{mcs}(w)$.

Finally, the *canonical model for* Δ_0 is defined as the structure $\mathcal{M}^c = (W^c, \mathsf{LEG}^c, \mathsf{ILL}^c, I^c, V^c)$.

Lemma A.5 *For any $w \in W^c$ and any δ and δ', $\mathcal{M}^c, w \models \mathbf{P}(\delta \mid \delta')$ iff $\mathsf{mcs}(w) \vdash \mathbf{P}(\delta \mid \delta')$.*

Proof. First, without loss of generality, we can assume association to the right if δ and δ' are sequences of atomic actions.

Suppose $\mathcal{M}^c, w \not\models \mathbf{P}(\delta \mid \delta')$. Then there is $\sigma \in I^c(\delta)$ and $\sigma' \in I^c(\delta')$ such that $\sigma[1] = w$, $\sigma'[1] = \sigma[f]$, and $(\sigma, \sigma') \notin \mathsf{LEG}^c$. By construction of LEG^c, $\mathsf{mcs}(w) \nvdash \mathbf{P}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma'))$. By Lemma A.3, $\mathsf{trace}(\sigma) = \delta$ and $\mathsf{trace}(\sigma') = \delta'$. Hence, $\mathsf{mcs}(w) \nvdash \mathbf{P}(\delta \mid \delta')$.

Suppose $\mathsf{mcs}(w) \nvdash \mathbf{P}(\delta \mid \delta')$. We have to consider four different cases: (i) $\delta = \delta' = \mathbf{skip}$; (ii) $\delta = \mathbf{skip}$ and δ' is a sequence of atomic actions; (iii) δ is a sequence of atomic actions and $\delta' = \mathbf{skip}$; (iv) both δ and δ' are sequences of atomic actions. Here we consider cases (i) and (iv).

For case (i), assume that $\delta = \delta' = \mathbf{skip}$. Since $\mathsf{trace}(w) = \mathbf{skip}$, it holds that $\mathsf{mcs}(w) \nvdash \mathbf{P}(\mathsf{trace}(w) \mid \mathsf{trace}(w))$. By the construction of LEG^c it holds that $(w, w) \notin \mathsf{LEG}^c$. Then clearly $\mathcal{M}^c, w \not\models \mathbf{P}(\mathbf{skip} \mid \mathbf{skip})$.

For case (iv), assume that $\delta = a_1; \ldots; a_k$ and $\delta' = b_1; \ldots; b_l$. Since $\mathsf{mcs}(w)$ is maximally consistent, $\mathsf{mcs}(w) \vdash \neg\mathbf{P}(\delta \mid \delta')$, so by Axiom (D3), $\mathsf{mcs}(w) \vdash \langle \delta; \delta' \rangle \top$. By the standard reduction axioms for dynamic modalities, $\mathsf{mcs}(w) \vdash \langle a_1 \rangle \ldots \langle a_k \rangle \langle b_1 \rangle \ldots \langle b_l \rangle \top$. By Lemma A.4 there is $\sigma \in I^c(\delta; \delta')$ such that $\sigma[1] = w$. Then $\sigma = \tau_1 \circ \tau_2$, with $\tau_1 \in I^c(\delta)$ and $\tau_2 \in I^c(\delta')$. By Lemma A.3, $\mathsf{trace}(\tau_1) = \delta$ and $\mathsf{trace}(\tau_2) = \delta'$. Then $\mathsf{mcs}(w) \nvdash \mathbf{P}(\mathsf{trace}(\tau_1) \mid \mathsf{trace}(\tau_2))$, so by construction of LEG^c, $(\tau_1, \tau_2) \notin \mathsf{LEG}^c$. It follows that $\mathcal{M}^c, w \not\models \mathbf{P}(\delta \mid \delta')$. □

Lemma A.6 *For any $w \in W^c$ and any δ and δ', $\mathcal{M}^c, w \models \mathbf{F}(\delta \mid \delta')$ iff $\mathsf{mcs}(w) \vdash \mathbf{F}(\delta \mid \delta')$.*

Proof. As in the proof of the previous lemma, we assume association to the right for δ and δ'.

Suppose $\mathcal{M}^c, w \models \mathbf{F}(\delta \mid \delta')$. Then there is $\sigma \in I^c(\delta)$ and $\sigma' \in I^c(\delta')$ such that $\sigma[1] = w$, $\sigma'[1] = \sigma[f]$ and $(\sigma, \sigma') \in \mathsf{ILL}^c$. By construction of ILL^c, $\mathsf{mcs}(w) \vdash \mathbf{F}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma'))$. By Lemma A.3, $\mathsf{trace}(\sigma) = \delta$ and $\mathsf{trace}(\sigma') = \delta'$, so $\mathsf{mcs}(w) \vdash \mathbf{F}(\delta \mid \delta')$.

Suppose $\mathsf{mcs}(w) \vdash \mathbf{F}(\delta \mid \delta')$. As in the previous lemma, there are four different possibilities for δ and δ' to consider: here, we consider only the case where both δ and δ' are sequences of atomic actions. Assume that δ and δ' are sequences of atomic actions. By Axiom (D6), $\mathsf{mcs}(w) \vdash \langle \delta; \delta' \rangle \top$. By Lemma A.4, there are paths $\tau_1 \in I^c(\delta)$ and $\tau_2 \in I^c(\delta')$ such that $\tau_1[1] = w$ and $\tau_2[1] = \tau_1[f]$. By Lemma A.3, $\mathsf{trace}(\tau_1) = \delta$ and $\mathsf{trace}(\tau_2) = \delta'$. Then $\mathsf{mcs}(w) \vdash \mathbf{F}(\mathsf{trace}(\tau_1) \mid \mathsf{trace}(\tau_2))$, so by construction of ILL^c, $(\tau_1, \tau_2) \in \mathsf{ILL}^c$. Then $\mathcal{M}^c, w \models \mathbf{F}(\delta \mid \delta')$. □

Lemma A.7 *For any $w \in W^c$ and any $\varphi \in \mathcal{L}_{red}$, $\mathcal{M}^c, w \models \varphi$ iff $\mathsf{mcs}(w) \vdash \varphi$.*

Finally, we are ready to prove the completeness part of Theorem 2.8, i.e. that for any $\Phi \subseteq \mathcal{L}$ and any $\psi \in \mathcal{L}$, if $\Phi \models \psi$ then $\Phi \vdash \psi$.

Proof of completeness part of Theorem 2.8 Suppose $\Phi \nvdash \psi$. By Lemma A.2, $t(\Phi) \nvdash t(\psi)$, where t is the translation referred to in Lemma A.2, and $t(\Phi) = \{t(\varphi) \mid \varphi \in \Phi\}$. Then $t(\Phi) \cup \{\neg t(\varphi)\}$ is consistent. Let Δ_0 be a maximally consistent extension of $t(\Phi) \cup \{\neg t(\varphi)\}$ and let $\mathcal{M}^c$ be the canonical model for Δ_0. Then $\mathsf{mcs}(w_0) \vdash t(\varphi)$ for all $t(\varphi) \in t(\Phi)$, and $\mathsf{mcs}(w_0) \nvdash t(\psi)$. By Lemma A.7, $\mathcal{M}^c, w_0 \models t(\varphi)$ for all $t(\varphi) \in t(\Phi)$, and $\mathcal{M}^c, w_0 \not\models t(\psi)$. Then $t(\Phi) \not\models t(\psi)$. By Lemma A.2 and soundness, $\Phi \not\models \psi$. □

B Corresponding validities

Proof of Theorem 3.1 Here, we provide an illustrative proof of one of the items in Theorem 3.1: LEG is closed under **FoMo** iff $\mathcal{F} \models \mathbf{P}(a \mid b; c) \to \mathbf{P}(a; b \mid c)$. The other items are proved using similar arguments.

Suppose $\mathcal{F} \not\models \mathbf{P}(a \mid b; c) \to \mathbf{P}(a; b \mid c)$. Then there is a model $\mathcal{M}$ based on $\mathcal{F}$ and a state w such that (i) $\mathcal{M}, w \models \mathbf{P}(a \mid b; c)$, but (ii) $\mathcal{M}, w \not\models \mathbf{P}(a; b \mid c)$. From (i), it follows that for all $\sigma \in I(a)$ and all $\sigma' \in I(b; c)$, if $\sigma[1] = w$ and $\sigma'[1] = \sigma[f]$, then $(\sigma, \sigma') \in \mathsf{LEG}$. From (ii), it follows that there are paths τ_1, τ_2 and τ_3 such that $\tau_1 \in I(a)$, $\tau_2 \in I(b)$, $\tau_3 \in I(c)$, $(\tau_1 \circ \tau_2)[1] = w$ and $\tau_3[1] = (\tau_1 \circ \tau_2)[f]$, and $(\tau_1 \circ \tau_2, \tau_3) \notin \mathsf{LEG}$. Since $\tau_1 \circ \tau_2 \in I(a; b)$ and $\tau_3 \in I(c)$, it holds that $(\tau_1, \tau_2 \circ \tau_3) \in \mathsf{LEG}$. Thus, LEG is not closed under **FoMo**.

Suppose $\mathcal{F} \models \mathbf{P}(a \mid b; c) \to \mathbf{P}(a; b \mid c)$. Take any paths τ_1, τ_2 and τ_3 such that $(\tau_1, \tau_2 \circ \tau_3) \in \mathsf{LEG}$. Construct the model $\mathcal{M}$ based on $\mathcal{F}$ such that $I(a) = \{\tau_1\}$, $I(b) = \{\tau_2\}$ and $I(c) = \{\tau_3\}$. Let $\tau_1[1] = w$. Then $\mathcal{M}, w \models \mathbf{P}(a \mid b; c)$, so by the initial assumption it holds that $\mathcal{M}, w \models \mathbf{P}(a; b \mid c)$. Since $I(a) = \{\tau_1\}$, $I(b) = \{\tau_2\}$ and $\tau_2[1] = \tau_1[f]$, it holds that $\tau_1 \circ \tau_2 \in I(a; b)$. Since $I(c) = \{\tau_3\}$ and $\tau_3[1] = (\tau_1 \circ \tau_2)[f]$, it follows that $(\tau_1 \circ \tau_2, \tau_3) \in \mathsf{LEG}$. Since τ_1, τ_2 and τ_3 were chosen arbitrarily, it follows that LEG is closed under **FoMo**. □

C Completeness proof for extended logics

Proof of completeness part of Theorem 3.2 The proof follows the same structure as the proof of the completeness part of Theorem 2.8 in Appendix A, but we also have to show that the relations LEG^c and ILL^c in the canonical model satisfy the properties corresponding to the added axioms. Here, we prove some illustrative examples:

- Assume that the formula corresponding to **NoCon**, i.e. $\mathbf{P}(a \mid b) \rightarrow \neg\mathbf{F}(a \mid b)$, is added as an axiom. Assume that $(\sigma, \sigma') \in \mathsf{LEG}^c$. By the definition of LEG^c, it holds that $\mathsf{mcs}(\sigma[1]) \vdash \mathbf{P}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma'))$. By the axiom $\mathbf{P}(a \mid b) \rightarrow \neg\mathbf{F}(a \mid b)$ and the inference rule of *uniform action substitution*, $\mathsf{mcs}(\sigma[1]) \vdash \neg\mathbf{F}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma'))$. Since $\mathsf{mcs}(\sigma[1])$ is consistent, it holds that $\mathsf{mcs}(\sigma[1]) \not\vdash \mathbf{F}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma'))$. Then $(\sigma, \sigma') \notin \mathsf{ILL}^c$ by the definition of ILL^c. Hence, LEG^c and ILL^c together satisfy **NoCon**.
- Assume that the formula corresponding to **NoGap**, i.e. $\neg\mathbf{F}(a \mid b) \rightarrow \mathbf{P}(a \mid b)$, is added as an axiom. Assume that $(\sigma, \sigma') \notin \mathsf{ILL}^c$ and that $\sigma'[1] = \sigma[f]$. Then $\mathsf{mcs}(\sigma[1]) \not\vdash \mathbf{F}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma'))$ by the definition of ILL^c. Since $\mathsf{mcs}(\sigma[1])$ is maximally consistent, it holds that $\mathsf{mcs}(\sigma[1]) \vdash \neg\mathbf{F}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma'))$. By the axiom $\neg\mathbf{F}(a \mid b) \rightarrow \mathbf{P}(a \mid b)$ and *uniform action substitution*, $\mathsf{mcs}(\sigma[1]) \vdash \mathbf{P}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma'))$, so by the definition of LEG^c, $(\sigma, \sigma') \in \mathsf{LEG}^c$. Hence, LEG^c and ILL^c together satisfy **NoGap**.
- Assume that the formula corresponding to **FoMo**, i.e. $\mathbf{P}(a \mid b; c) \rightarrow \mathbf{P}(a; b \mid c)$, is added as an axiom. Assume that $(\sigma, \sigma' \circ \sigma'') \in \mathsf{LEG}^c$. By the definition of LEG^c, $\mathsf{mcs}(\sigma[1]) \vdash \mathbf{P}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma' \circ \sigma''))$. Then, $\mathsf{mcs}(\sigma[1]) \vdash \mathbf{P}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma'); \mathsf{trace}(\sigma''))$. By the axiom $\mathbf{P}(a \mid b; c) \rightarrow \mathbf{P}(a; b \mid c)$ and *uniform action substitution*, we can get $\mathsf{mcs}(\sigma[1]) \vdash \mathbf{P}(\mathsf{trace}(\sigma); \mathsf{trace}(\sigma') \mid \mathsf{trace}(\sigma''))$. Then $\mathsf{mcs}(\sigma[1]) \vdash \mathbf{P}(\mathsf{trace}(\sigma \circ \sigma') \mid \mathsf{trace}(\sigma''))$. By the definition of LEG^c, $(\sigma \circ \sigma', \sigma'') \in \mathsf{LEG}^c$. Hence, LEG^c is closed under **FoMo**.
- Assume that the formula corresponding to **BaMo**, i.e. $\mathbf{F}(a; b \mid c) \rightarrow \mathbf{F}(a \mid b; c)$, is added as an axiom. Assume that $(\sigma \circ \sigma', \sigma'') \in \mathsf{ILL}^c$. By the definition of ILL^c, $\mathsf{mcs}(\sigma[1]) \vdash \mathbf{F}(\mathsf{trace}(\sigma \circ \sigma') \mid \mathsf{trace}(\sigma''))$. Then, $\mathsf{mcs}(\sigma[1]) \vdash \mathbf{F}(\mathsf{trace}(\sigma); \mathsf{trace}(\sigma') \mid \mathsf{trace}(\sigma''))$. By the axiom $\mathbf{F}(a; b \mid c) \rightarrow \mathbf{F}(a \mid b; c)$, and *uniform action substitution*, we know $\mathsf{mcs}(\sigma[1]) \vdash \mathbf{F}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma'); \mathsf{trace}(\sigma''))$. Then, $\mathsf{mcs}(\sigma[1]) \vdash \mathbf{F}(\mathsf{trace}(\sigma) \mid \mathsf{trace}(\sigma' \circ \sigma''))$. By the definition of ILL^c, $(\sigma, \sigma' \circ \sigma'') \in \mathsf{ILL}^c$. Hence, ILL^c is closed under **BaMo**.

□

References

[1] Ågotnes, T., W. van der Hoek, J. Rodriguéz-Aguilar, C. Sierra and M. Wooldridge, *A temporal logic of normative systems*, in: D. Makinson, J. Malinowski and H. Wansing,

editors, *Towards Mathematical Philosophy. Trends in Logic, vol 28.*, Springer, Dordrecht, 2009 pp. 69–106.

[2] Anglberger, A. J. J., *Dynamic deontic logic and its paradoxes*, Studia Logica **80** (2008), pp. 427–435.

[3] Arntzenius, F., A. Elga and J. Hawthorne, *Bayesianism, infinite decisions, and binding*, Mind **113** (2004), pp. 251–283.

[4] Bales, A. and C. Benn, *Supererogation and sequence*, Synthese **198** (2021), pp. 7763–7780.

[5] Broersen, J., *Action negation and alternative reductions for dynamic deontic logics*, Journal of Applied Logic **2** (2004), pp. 153–168.

[6] Broersen, J., R. Wieringa and J.-J. Meyer, *A fixed-point characterization of a deontic logic of regular action*, Fundamenta Informaticae **48** (2001), pp. 3–107.

[7] Ju, F., K. Nygren and T. Xu, *Modeling legal conflict resolution based on dynamic logic*, Journal of Logic and Computation **31** (2021), pp. 1102–1128.

[8] Kamp, H., *Free choice permission*, Proceedings of the Aristotelian Society **74** (1973), pp. 57–74.

[9] Kulicki, P. and R. Trypuz, *Completely and partially executable sequences of actions in deontic context*, Synthese **192** (2015), pp. 1117–1138.

[10] Meyer, J.-J. C., *A different approach to deontic logic: deontic logic viewed as a variant of dynamic logic*, Notre Dame Journal of Formal Logic **29** (1988), pp. 109–136.

[11] Prisacariu, C. and G. Schneider, *A dynamic deontic logic for complex contracts*, The Journal of Logic and Algebraic Programming **81** (2012), pp. 458–490.

[12] Schmidt, R. A. and D. Tishkovsky, *On combinations of propositional dynamic logic and doxastic modal logics*, Journal of Logic, Language and Information **17** (2008), pp. 109–129.

[13] Segerberg, K., *Blueprint for a dynamic deontic logic*, Journal of Applied Logic **7** (2009), p. 388–402.

[14] van der Meyden, R., *The dynamic logic of permission*, Journal of Logic and Computation **6** (1996), pp. 465–479.

[15] von Wright, G. H., “Norm and Action. A Logical Enquiry,” Routledge and Kegan Paul, London, 1963.

[16] von Wright, G. H., *Deontic logic: A personal view*, Ratio Juris **12** (1999), pp. 26–38.

An Andersonian-Kangerian Reduction of Term-Modal Deontic Logics

Stef Frijters [1]

Institute of Philosophy (HIW)
Kardinaal Mercierplein 2
3000 Leuven / Belgium

Abstract

In the 17th century, Leibniz analyzed obligatoriness as "what is necessary for a good person to do". Almost 300 years later, Anderson and Kanger independently came to a similar analysis. Both proposed alethic modal logics with a deontic constant, in which an obligation operator can be defined. It can be proven that several deontic logics are fragments of these alethic modal logics. Åqvist calls this translation theorem, "one of the main mathematical results on propositional monadic deontic logic". We will show that a similar translation theorem can be proven for a number of predicative deontic logics, known as term-modal deontic logics (TMDLs). TMDLs were recently developed and allow one to explicitly represent quantification over bearers and counterparties of obligations. In doing so, they avoid the pitfalls of earlier attempts at developing predicative deontic logics.
In this paper we define several TMDLs and several alethic modal logics with predicative constants, as well as a translation from the TMDLs to the alethic modal logics. We show that this translation is actually closer to the original Leibnizian analysis of obligation than what is provided in the standard Andersonian-Kangerian systems. We also prove that each of the TMDLs is a fragment of one of the alethic modal logics. This is interesting not only from a conceptual, but also from a technical perspective. It shows that (some) term-modal logics are actually reducible to standard (non term-modal) modal logics.

Keywords: Deontic logic, term-modal logic, Andersonian-Kangerian reduction

1 Introduction

The best studied system of deontic logic is Standard Deontic Logic (**SDL**). **SDL** is just the normal propositional modal logic **KD**, where the modal oper-

[1] stef.frijters@kuleuven.be This research was funded by the research project 'BITSHARE: Bitstring Semantics for Human and Artificial Reasoning' (IDN-19-009, Internal Funds KU Leuven) and by the Research Foundation - Flanders, through the project "Towards a more integrated formal account of actual ethical reasoning, with applications in medical ethics." (G0D2716N). The author would like to thank Frederik Van De Putte, Lorenz Demey, Jan Heylen, Robbe Van Den Eede and the anonymous reviewers for their valuable feedback on earlier versions of this paper.

ator, written as O, is interpreted as obligation. A different approach to deontic logic was introduced by Alan Ross Anderson [1,2] and Stig Kanger [15,13], although the philosophical idea behind this approach can be traced back to the philosopher and mathematician G. W. Leibniz. The latter analysed 'φ is obligatory for a' as 'φ is necessary for a's being a good person' [14]. Inspired by a similar intuition, Anderson and Kanger do not use a primitive obligation operator, but instead an alethic modal operator and a propositional constant. They then define the obligation operator in this new language.

Kanger and Anderson work this out in different ways. In the Kangerian approach, the propositional constant G is to be read as "what morality prescribes" [15]. It is obligatory that φ, denoted as $\mathsf{O}\varphi$, is then defined as "it is necessary for what morality prescribes that φ", $\Box(\mathsf{G} \rightarrow \varphi)$. In the Andersonian approach, a different propositional constant, S, is used. This denotes some bad state of affairs [1, p. 103], a violation of a rule [2, p. 348] or a sanction being applicable [14]. It is obligatory that φ, also denoted as $\mathsf{O}\varphi$, can then be formally defined as $\Box(\neg\varphi \rightarrow \mathsf{S})$. This can be read as 'not φ necessarily implies a bad state of affairs', 'not φ necessarily implies a violation of a rule' or 'not φ necessarily implies the applicability of a sanction'. These two approaches are equivalent if we simply take S to be the negation of G [18].

It turns out that all logical principles validated by **SDL** and other logics like it, are also validated by Andersonian-Kangerian logics. One can define a translation that assigns to each formula of **SDL** (or a closely related system) a formula of an Andersonian-Kangerian logic. It has been proven that for every formula φ, φ is a theorem of **SDL** iff the translation of φ is a theorem of the Andersonian-Kangerian logic. Åqvist calls this translation theorem "one of the main mathematical results on propositional monadic deontic logic" [4]. In this paper, our main goal is to expand this result to *predicative* deontic logic.

Predicative (or first-order) deontic logic has long been understudied. Presumably, this was because it was either thought to be trivial, in view of the results on first-order modal logic in general, or because there are certain problems (e.g. the interpretation of the Barcan Formula in a deontic context) for which there was no solution at hand [14,7]. Whatever the reason might have been, the result is that most systems of deontic logic cannot capture elementary patterns of deontic reasoning that involve quantification.

Recently, this has changed with the introduction of term-modal deontic logics (**TMDLs**) [7,9,19]. The language of these **TMDLs** contains obligation operators indexed with terms (variables or constants) of the language. This allows for a great increase in expressivity, as illustrated in Table 1 (taken from [7,9]). In addition, this indexing solves the problems plaguing earlier attempts at constructing a system of first-order deontic logic (as argued in [7,9]).

In this paper, we prove a translation theorem for **TMDLs**. We define a family of logics that model the Andersonian-Kangerian ideas in a predicative context. These logics use a standard (that is, not term-modal) modal operator and two predicative constants $\mathcal{Q}$ and $\mathcal{Q}^2$. So for example, 'it is obligatory for a that φ', $\mathsf{O}_a\varphi$, can be defined in these logics as 'a being a good person

It is obligatory for Alfred that φ	$\mathsf{O}_a\varphi$
φ is obligatory for every philosopher	$(\forall x)(Px \to \mathsf{O}_x\varphi)$
Everyone has an obligation towards b to help b	$(\forall x)(\mathsf{O}_x^b Hxb)$
b has an obligation towards everyone else to help them	$(\forall x)(x \neq b \to \mathsf{O}_b^x Hbx)$
Every parent of a sick child has an obligation towards the child to care for it	$(\forall x)(\forall y)((Pxy \wedge Sy) \to \mathsf{O}_x^y Cxy)$

Table 1
Illustrating the expressiveness of **TMDL**

necessarily implies φ’, $\Box(\mathcal{Q}a \to \varphi)$. [2]

The paper is structured as follows. In Section 2 we define and discuss a family of **TMDLs**, and in Section 3 we do the same for a family of first-order modal logics based on the Andersonian-Kangerian ideas. Section 4 is dedicated to proving the translation theorem for each of these logics. Finally, Section 5 summarizes the results, briefly discusses their implications, and sketches possible paths for future research.

2 Term-modal deontic logics

Term-modal deontic logics are based on term-modal logics that were themselves developed for epistemic logic [6,16]. The main innovation is that instead of a standard modal operator, the language of term-modal logic contains a modal operator that is indexed by a term (variable or constant) of the language. Semantically, this is mirrored by the use of a ternary (or quaternary) instead of a binary accessibility relation.

In this section we start out by defining the language that is shared by all **TMDLs** discussed in this paper (Section 2.1). We then define the semantics (Section 2.2) and an axiomatisation (Section 2.3) of the weakest **TMDL** that we discuss in this paper: **TMK**. This is a term-modal variant of the normal modal logic **K**. In Section 2.4 we define different extensions of **TMK**.

2.1 The language of TMDLs

The definition of the language of **TMDLs** is taken from previous work [7,9]. Let $C = \{a, b, \ldots\}$ be the set of constants and $V = \{x, y, \ldots\}$ be the set of variables. We let $\alpha, \beta, \ldots$ range over C and $\nu, \xi, \ldots$ over V. Let $T = C \cup V$ be the set of terms (always denoting persons) and $\theta, \kappa, \ldots$ the metavariables ranging over it. For each natural number n we let $\mathcal{P}^n$ be a set of n-ary predicate symbols and we let $\mathcal{P}$ be the union of all $\mathcal{P}^n$. We let P range over $\mathcal{P}$. Lastly, we let φ, ψ, χ be metavariables for formulas and we use Γ, Δ, Θ as metavariables for sets of formulas. Our language $\mathcal{L}$ is defined by the following Backus-Naur

[2] In chapter four of the PhD-thesis [7] such a translation theorem is given for one specific logic. In the terminology of this paper that is the logic **TML** extended with all of the axiom schemes in Table 4.

form:

$$\varphi ::= P\theta_1 \ldots \theta_n \mid \theta = \kappa \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathsf{O}_\theta\varphi \mid \mathsf{O}^\theta_\kappa\varphi \mid (\forall\nu)\varphi$$

The other Boolean connectives are defined in the standard way. Additionally, $(\exists\nu)\varphi =_{\mathsf{df}} \neg(\forall\nu)\neg\varphi$, $\mathsf{P}_\theta\varphi =_{\mathsf{df}} \neg\mathsf{O}_\theta\neg\varphi$ and $\mathsf{P}^\theta_\kappa\varphi =_{\mathsf{df}} \neg\mathsf{O}^\theta_\kappa\neg\varphi$. We will often write $\theta \neq \kappa$ instead of $\neg\theta = \kappa$.

The notion of free and bound variables are as usual, with two additions (cfr. Fitting et al. [6]): (1) The free occurrences of variables in $\mathsf{O}_\theta\varphi$ are all free occurrences of variables in φ and in addition θ if θ is a variable, and (2) the free occurrences of variables in $\mathsf{O}^\kappa_\theta\varphi$ are θ, if θ is a variable, κ, if κ is a variable, and all free occurrences of variables in φ. We define the set $\mathcal{S}$ of sentences as consisting of those formulas $\varphi \in \mathcal{L}$ such that all variables in φ are bound.

The two term-modal operators O_θ and O^κ_θ denote undirected and directed personal obligations. A formula $\mathsf{O}_\theta\varphi$ is read as 'θ has an obligation that φ' and $\mathsf{O}^\kappa_\theta\varphi$ as 'θ has an obligation towards κ that φ'. Here, θ is the bearer of the obligation, i.e. the person who is at fault if the obligation has been violated, and κ is the counterparty of the obligation.

The second term-modal operator (O^κ_θ) is very useful to model reasoning with the Hohfeldian rights-relations. For example, 'a has the right to vote' becomes in the Hohfeldian analysis: 'everyone else has towards a the obligation not to prevent a from voting'. This can be formalised as: $(\forall x)((x \neq a) \to \mathsf{O}^a_x\neg Pxa)$, where Pxy is interpreted as 'x prevents y from voting'. The expressiveness of this language is further illustrated in Table 1. For a more detailed discussion of this language, see [7,9].

2.2 Semantics of TMK

TMK-models (Definition 2.1) do not differ substantially from other models for first-order modal logic with constant domains, except for the two accessibility relations.

Definition 2.1 A **TMK**-model is a tuple $M = \langle W, \mathcal{A}, R, R^D, I\rangle$, where:

1. $W \neq \emptyset$
2. $\mathcal{A} \neq \emptyset$
3. $R \subseteq W \times \mathcal{A} \times W$
4. $R^D \subseteq W \times \mathcal{A} \times \mathcal{A} \times W$
5. I is an *interpretation* function that assigns to every $\theta \in T$ a $p \in \mathcal{A}$ and to every pair $\langle P, w\rangle \in \mathcal{P}^n \times W$ an element of $\wp(\mathcal{A}^n)$ for every natural number $n \in \mathbb{N}$.

The set W is the world domain, consisting of possible worlds $w, w_1, \ldots$ and $\mathcal{A}$ is the agent domain, consisting of agents $p, p_1, p_2, \ldots$ Both are non-empty and are allowed to be at most countably infinite.[3] R and R^D are two accessibility relations. The set $R(w, p_1) =_{\mathsf{df}} \{w' \in W \mid \langle w, p_1, w'\rangle \in R\}$ is interpreted as

[3] The agent domain is the analogue of the (object) domain of first-order logic. It is possible to expand the agent domain to include objects that are not agents. This is discussed in Section 3.5.3 of [7].

the set of worlds where all the obligations that agent p_1 has in world w have been fulfilled. Similarly, $R(w, p_1, p_2) =_{\mathsf{df}} \{w' \in W \mid \langle w, p_1, p_2, w'\rangle \in R^D\}$ is interpreted as the set of worlds where all obligations that agent p_1 has towards p_2 in world w have been fulfilled.

After definining what it means to be a ν-alternative (Definition 2.2), we are in a position to give the semantic clauses (Definition 2.3). The only non-standard clauses are SC5 and SC6. Semantic consequence and validity are defined in the usual way (Definitions 2.4 and 2.5). In what follows, we will omit the subscript **TMK** when this is clear from the context.

Definition 2.2 For any $\nu \in V$, $M' = \langle W, \mathcal{A}, R, R^D, I'\rangle$ is a ν-alternative to $M = \langle W, \mathcal{A}, R, R^D, I\rangle$ iff I' differs at most from I in the member of $\mathcal{A}$ that I' assigns to ν.

Definition 2.3 [Semantic Clauses] For any **TMK**-model $M = \langle W, \mathcal{A}, R, R^D, I\rangle$:
SC1 $M, w \models P\theta_1 \ldots \theta_n$ iff $\langle I(\theta_1), \ldots, I(\theta_n)\rangle \in I(P, w)$
SC2 $M, w \models \neg\varphi$ iff $M, w \not\models \varphi$
SC3 $M, w \models \varphi \vee \psi$ iff $M, w \models \varphi$ or $M, w \models \psi$
SC4 $M, w \models \theta = \kappa$ iff $I(\theta) = I(\kappa)$
SC5 $M, w \models \mathsf{O}_\theta\varphi$ iff $M, w' \models \varphi$ for all $w' \in R(w, I(\theta))$
SC6 $M, w \models \mathsf{O}_\theta^\kappa\varphi$ iff $M, w' \models \varphi$ for all $w' \in R^D(w, I(\theta), I(\kappa))$
SC7 $M, w \models (\forall\nu)\varphi$ iff for every ν-alternative M': $M', w \models \varphi$

Definition 2.4 [Semantic consequence] φ is a semantic consequence of Γ, $\Gamma \Vdash_{\mathbf{TMK}} \varphi$ iff for every **TMK**-model $M = \langle W, \mathcal{A}, R, R^D, I\rangle$ and $w \in W$: if $M, w \models \psi$ for all $\psi \in \Gamma$, then $M, w \models \varphi$.

Definition 2.5 [Validity] φ is valid, $\Vdash_{\mathbf{TMK}} \varphi$ iff for every **TMK**-model $M = \langle W, \mathcal{A}, R, R^D, I\rangle$ and $w \in W$: $M, w \models \varphi$.

2.3 Axiomatisation of TMK

A sound and strongly complete axiomatisation of **TMK** is obtained by closing a complete axiomatisation of classical propositional logic (**CL**) with all instances of the axiom schemata in Table 2 under the rules of Table 3.[4] $\varphi(\theta/\kappa)$ is the result of replacing all free occurrences of κ in φ by θ, relettering bound variables if necessary to avoid rendering new occurrences of θ bound in $\varphi(\theta/\kappa)$. $\varphi(\theta//\kappa)$ is the result of replacing various (not necessarily all or even any) free occurrences of θ in φ by occurrences of κ, again relettering if necessary [20].

Theoremhood and derivability are defined as follows. A formula φ is a **TMK**-theorem (denoted $\vdash_{\mathbf{TMK}} \varphi$) iff φ can be derived from the **TMK**-axioms and rules. $\varphi \in \mathcal{S}$ is **TMK**-derivable from $\Gamma \subseteq \mathcal{S}$ (denoted $\Gamma \vdash_{\mathbf{TMK}} \varphi$) iff there are $\psi_1, \ldots, \psi_n \in \Gamma$ such that $\vdash_{\mathbf{TMK}} (\psi_1 \wedge \ldots \wedge \psi_n) \to \varphi$. From this it follows immediately that $\vdash_{\mathbf{TMK}}$ is compact. We write $\vdash$ instead of $\vdash_{\mathbf{TMK}}$ where this does not lead to confusion.

[4] The proof of soundness and strong completeness is a straightforward variation on the completeness proof for **TMDL** in [7,9].

(K)	$\mathsf{O}_\alpha(\varphi \to \psi) \to (\mathsf{O}_\alpha\varphi \to \mathsf{O}_\alpha\psi)$	(UI)	$(\forall\nu)\varphi \to \varphi(\alpha/\nu)$
(BF)	$(\forall\nu)\mathsf{O}_\alpha\varphi \to \mathsf{O}_\alpha(\forall\nu)\varphi$	(REF)	$\alpha = \alpha$
(DK)	$\mathsf{O}^\beta_\alpha(\varphi \to \psi) \to (\mathsf{O}^\beta_\alpha\varphi \to \mathsf{O}^\beta_\alpha\psi)$	(SUB)	$(\alpha = \beta) \to (\varphi \to \varphi(\alpha//\beta))$
(DBF)	$(\forall\nu)\mathsf{O}^\beta_\alpha\varphi \to \mathsf{O}^\beta_\alpha(\forall\nu)\varphi$	(ND)	$(\alpha \neq \beta) \to \mathsf{O}_\gamma(\alpha \neq \beta)$
		(DND)	$(\alpha \neq \beta) \to \mathsf{O}^\delta_\gamma(\alpha \neq \beta)$

Table 2
Axiom schemata of **TMK**

(MP)	if $\varphi \to \psi$ and φ, then ψ
(NEC)	if $\vdash \varphi$, then $\vdash \mathsf{O}_\alpha\varphi$
(DNEC)	if $\vdash \varphi$, then $\vdash \mathsf{O}^\beta_\alpha\varphi$.
(UG)	if $\vdash \varphi \to \psi(\alpha/\nu)$ and α not in φ or ψ, then $\vdash \varphi \to (\forall\nu)\psi$.

Table 3
Rules of **TMK**

2.4 Extensions of TMK

One of the points of contention in propositional deontic logic is what principles should be accepted. For example, the scheme $\mathsf{O}\varphi \to \neg\mathsf{O}\neg\varphi$ is valid in **SDL**, but in the face of deontic conflicts one could argue that this should be given up [12]. These discussions are also relevant for **TMDLs**: **TMK** does not validate $\mathsf{O}_\theta\varphi \to \neg\mathsf{O}_\theta\neg\varphi$, but an extension of the logic might. We will not take a stand on such discussions here. Instead, we propose a family of **TMDLs**. Each member of this family is obtained by adding different axiom schemes and corresponding frame conditions from Table 4 to **TMK**. For example, the **TMDL** discussed in [9] is obtained by adding the axiom schemes (D) and (DIU) (and the corresponding frame conditions) to **TMK**.[5] For all these logics, the proof of soundness and strong completeness is a straightforward variation on the completeness proof for **TMDL** in [7,9].

(D)	$\mathsf{O}_\theta\varphi \to \mathsf{P}_\theta\varphi$	For all $p \in \mathcal{A}$ and $w \in W$, $R(w,p) \neq \emptyset$
(DD)	$\mathsf{O}^\kappa_\theta\varphi \to \mathsf{P}^\kappa_\theta\varphi$	For all $p, p' \in \mathcal{A}$ and $w \in W$, $R^D(w,p,p') \neq \emptyset$
(+)	$\mathsf{O}_\theta(\mathsf{O}_\theta\varphi \to \varphi)$	For all $w, w' \in W$ and $p \in \mathcal{A}$, $R(w,p) = R(w',p)$
(D+)	$\mathsf{O}^\kappa_\theta(\mathsf{O}^\kappa_\theta\varphi \to \varphi)$	For all $w, w' \in W$ and $p, p' \in \mathcal{A}$, $R^D(w,p,p') = R(w',p,p')$
(DIU)	$\mathsf{O}^\beta_\alpha\varphi \to \mathsf{O}_\alpha\varphi$	For all $w \in W$, $p_1, p_2 \in \mathcal{A}$: $R(w,p_1) \subseteq R(w,p_1,p_2)$

Table 4
Additional axiom schemes for **TMDLs**

Other axiom schemes could have been added to Table 4. However, we think that the schemes presented here are the most plausible candidates for inclusion in a **TMDL**. The schemes (D) and (DD) are the term-modal analogues of

[5] One could also argue that **TMK** itself is already too strong. We refer the interested reader to [10] for a discussion of term-modal logics weaker than **TMK**.

the famous ought-implies-can principle in propositional deontic logic. Under a plausible reading, the schemes (+) and (D+) can be used to model the intuition that iterated deontic modalities *of the same kind* do not add any meaning. For a deeper discussion of such intuitions for propositional deontic logic, see for example [5]. Finally, the scheme (DIU) concerns the interaction between directed and undirected personal obligations. Arguments for and against this principle in a predicative context were discussed in [7,9].[6]

3 The alethic modal logics

As mentioned in the introduction, the goal of this paper is to prove a translation theorem for different **TMDLs**. This means that we must also define a family of first-order logics inspired by Andersonian-Kangerian (and Leibnizian) intuitions. Each of these logics is built around an alethic modal operator and two predicative constants $\mathcal{Q}$ and $\mathcal{Q}^2$. To define this family of logics, we take the same approach as in Section 2. We first define the language shared by all the logics in this family in Section 3.1. Then we give the semantics and axiomatisation of the weakest logic in the family, **AK**, in Sections 3.2 and 3.3. Finally, Section 3.4 defines the extensions of **AK** by giving a list of axiom schemes (and corresponding frame conditions) that can be added to **AK**.

3.1 The language of AK

The language of **AK** is built using the same constants, variables and predicate symbols as the language of **TMDLs**.[7] We add to this a standard modal operator $\Box$, and two predicative constants $\mathcal{Q}$ and $\mathcal{Q}^2$ with arity 1 and 2 respectively. Our new language $\mathcal{L}^Q$ is defined by the following Backus-Naur form:

$$\varphi ::= P\theta_1 \ldots \theta_n \mid \theta = \kappa \mid \neg\varphi \mid \varphi \vee \varphi \mid \Box\varphi \mid (\forall\nu)\varphi \mid \mathcal{Q}\theta \mid \mathcal{Q}^2\theta\kappa$$

The other connectives are defined in the standard way, with the addition that $\Diamond\varphi =_{\mathsf{df}} \neg\Box\neg\varphi$. The notions of free and bound variables are as usual. A wff φ is a sentence, $\varphi \in \mathcal{S}^Q$, iff all the variables in φ are bound. Note that the language of **AK** contains no term-modal operators, but only standard modal operators.

The main novelty in the language $\mathcal{L}^Q$ lies in the predicative constants. $\mathcal{Q}x$ can be read in different ways. Staying close to the Leibnizian intuitions, we will read $\mathcal{Q}x$ as 'x is a good person'. However, this choice is rather arbitrary. We can also stay closer to a Kangerian or Andersonian interpretation and read $\mathcal{Q}x$ as 'x's obligations have been fulfilled' or 'x is not in violation of any obligation'. From a technical point of view, nothing hinges on the precise interpretation of the constant. The same goes for $\mathcal{Q}^2$. The formula $\mathcal{Q}^2xy$ can be read as 'x is a good person when it comes to their dealings with y', 'x's obligations towards y have been fulfilled', or 'x is not in violation of any obligation that x has towards y. A predicate similar to $\mathcal{Q}^2$ was proposed by Lindahl [17, p. 163]

[6] A non-monotonic version of this principle was proposed in [8].

[7] This will be important for the proof of the translation theorem (Theorem 4.9).

to solve some problems with Kanger's formal treatment of rights. Following Lindahl's suggestion, we will read $\mathcal{Q}^2xy$ as 'y has not been wronged by x'.

3.2 Semantics of AK

The semantics of **AK** are fairly standard. Definition 3.1 gives the definition of **AK**-models. This does not differ significantly from that of familiar constant-domain models for first-order modal logic, except for the addition of the functions f and f^2. Definition 3.2 gives the semantic clauses for **AK**. A ν-alternative, semantic consequence and validity are all defined analogous to those definitions for **TMK** in Section 2.2.

Definition 3.1 An **AK**-model is a tuple $M = \langle W, \mathcal{A}, R^Q, f, f^2, I\rangle$, where:
1. $W \neq \emptyset$
2. $\mathcal{A} \neq \emptyset$
3. $R^Q \subseteq W \times W$
4. f and f^2 are functions such that:
4.1. $f : \mathcal{A} \to \wp(W)$
4.2. $f^2 : \mathcal{A} \times \mathcal{A} \to \wp(W)$
5. I is an *interpretation* function that assigns to every $\theta \in T$ a $p \in \mathcal{A}$ and to every pair $\langle P, w\rangle \in \mathcal{P}^n \times W$ an element of $\wp(\mathcal{A}^n)$ for every natural number $n \in \mathbb{N}$.

The relation R^Q should be given an alethic instead of a deontic interpretation: $R^Q(w) =_{\mathsf{df}} \{w' \mid \langle w, w'\rangle \in R^Q\}$ is the set of worlds that are possible at w. The function $f(p)$ can be interpreted as the set of worlds that are (deontically) ideal for agent p. Alternatively, we could say that $f(p)$ is the set of worlds where p is praiseworthy, where p is not blameworthy, or where p satisfies her obligations/the demands placed on her. Similarly, $f(p, p')$ is the set of worlds where p' has not been wronged by p, or where p's obligations toward p' have been fulfilled.

Definition 3.2 [Semantic Clauses] For any **AK**-model $M = \langle W, \mathcal{A}, R^Q, f, f^2, I\rangle$:
SC1 $M, w \models P\theta_1 \dots \theta_n$ iff $\langle I(\theta_1), \dots, I(\theta_n)\rangle \in I(P, w)$
SC2 $M, w \models \neg\varphi$ iff $M, w \not\models \varphi$
SC3 $M, w \models \varphi \vee \psi$ iff $M, w \models \varphi$ or $M, w \models \psi$
SC4 $M, w \models \theta = \kappa$ iff $I(\theta) = I(\kappa)$
SC5 $M, w \models \Box\varphi$ iff $M, w' \models \varphi$ for all $w' \in R^Q(w)$
SC6 $M, w \models (\forall\nu)\varphi$ iff for every ν-alternative M': $M', w \models \varphi$
SC7 $M, w \models \mathcal{Q}\theta$ iff $w \in f(I(\theta))$.
SC8 $M, w \models \mathcal{Q}^2\theta\kappa$ iff $w \in f^2(I(\theta), I(\kappa))$.

3.3 Axiomatisation of AK

A sound and strongly complete axiomatisation of **AK** is obtained by closing a complete axiomatisation of **CL** with all instances of the axiom schemata in Table 5 under the rules of Table 6. Theoremhood and derivability are defined in a manner completely analogous to that in Section 2.3. The proofs for soundness and strong completeness are safely left to the reader.

(QK)	$\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$	(QREF)	$\alpha = \alpha$
(QBF)	$(\forall\nu)\Box\varphi \to \Box(\forall\nu)\varphi$	(QSUB)	$(\alpha = \beta) \to (\varphi \to \varphi(\alpha//\beta))$
(QUI)	$(\forall\nu)\varphi \to \varphi(\alpha/\nu)$	(QND)	$(\alpha \neq \beta) \to \Box(\alpha \neq \beta)$

Table 5
Axiom schemata of **AK**

(QMP)	if $\varphi \to \psi$ and φ, then ψ
(QNEC)	if $\vdash \varphi$, then $\Box\varphi$
(QUG)	if $\vdash \varphi \to \psi(\alpha/\nu)$ and α not in φ or ψ, then $\vdash \varphi \to (\forall\nu)\psi$.

Table 6
Rules of **AK**

3.4 Extensions of AK

As we did for **TMK**, we will now define a family of extensions of **AK**. Each member of this family is obtained by adding any number of axiom schemes and corresponding frame conditions from Table 7 to **AK**. For all these logics, the proofs of soundness and strong completeness can safely be left to the reader.

(QQ)	$\mathcal{Q}\theta \to \mathcal{Q}^2\theta\kappa$	For all $p_1, p_2 \in \mathcal{A}$ and $w \in W$: $f(p_1) \subseteq f^2(p_1, p_2)$
(AQ)	$\Diamond\mathcal{Q}\theta$	For all $p \in \mathcal{A}$ and $w \in W$, $R^Q(w) \cap f(p) \neq \emptyset$
(ADQ)	$\Diamond\mathcal{Q}^2\theta\kappa$	For all $p, p' \in \mathcal{A}$ and $w \in W$, $R^Q(w) \cap f^2(p, p') \neq \emptyset$
(AT)	$\Box\varphi \to \varphi$	R is reflexive

Table 7
Additional axiom schemes for **AK**

The axiom schemes in Table 7 deserve a short explanation. The scheme (QQ) corresponds to the intuition that whenever one is a good person, then one has not wronged anyone. [8] This corresponds to the intuition behind (DIU). The schemes (AQ) and (ADQ) model the intuition that it is always possible to be a good person or not to wrong another person (i.e. to fulfill one's obligations to other persons). These schemes correspond to (D) and (DD). For an alethic interpretation of the $\Box$ operator, (AT) is usually considered to be a plausible axiom scheme: if φ is necessary, then φ is the case. We will need this scheme to prove the translation theorem for **TMDLs** that contain (+) and (D+). [9]

[8] One could even wonder about the other direction: if θ has not wronged anyone, then θ is a good person, $(\forall\nu)(\mathcal{Q}^2\theta\nu) \to \mathcal{Q}\theta$. Accepting this and (QQ) would mean that $\mathcal{Q}$ is definable in terms of $\mathcal{Q}^2$ (and vice versa).

[9] We could have added more schemes to the table. For example (A4) $\Box\varphi \to \Box\Box\varphi$ and (A5) $\Diamond\varphi \to \Box\Diamond\varphi$ are plausible candidate schemes for an alethic interpretation of the $\Box$ operator. However, these schemes play no role for the translation theorems in the next section.

4 The translation theorems

This section contains the most important technical result of this paper, the proof of the translation theorems for **TMDLs**. In Section 4.1 we define the translation from the language of **TMDLs** to that of the first-order Andersonian-Kangerian logics and comment on the implications of this translation. In Section 4.2 we prove the translation theorem for the two weakest logics, i.e. we show that **TMK** is the deontic fragment of **AK**, that for every φ that is valid in **TMK**, the translation of φ is valid in **AK**. Section 4.3 shows how this proof can be expanded to the extensions of **TMK** and **AK**.

4.1 The translation

Definition 4.1 gives a translation $\mathcal{T}$, based on [3, p. 114]. The translation is trivial except for clauses 5. and 6. These clauses capture the Leibnizian-Andersonian-Kangerian intuitions discussed in the introduction. In fact, we would argue that clause 5. better represents the Leibnizian analysis than the propositional proposals by Anderson and Kanger did. Leibniz is after all analyzing an (undirected) personal obligation with a bearer, which is a nuance that gets lost in the propositional approach.

Definition 4.1 [Translation] Let $\varphi \in \mathcal{L}$, then:

1. $\mathcal{T}(P\theta_1, \dots, \theta_n) = P\theta_1, \dots, \theta_n$
2. $\mathcal{T}(\theta = \kappa) = (\theta = \kappa)$
3. $\mathcal{T}(\neg\varphi) = \neg\mathcal{T}(\varphi)$
4. $\mathcal{T}(\varphi \vee \psi) = \mathcal{T}(\varphi) \vee \mathcal{T}(\psi)$
5. $\mathcal{T}(\mathsf{O}_\theta\varphi) = \Box(\mathcal{Q}\theta \to \mathcal{T}(\varphi))$
6. $\mathcal{T}(\mathsf{O}^{\kappa}_{\theta}\varphi) = \Box(\mathcal{Q}\theta\kappa \to \mathcal{T}(\varphi))$
7. $\mathcal{T}((\forall\nu)\varphi) = (\forall\nu)\mathcal{T}(\varphi)$

There is a difference in the expressiveness of the language of the **TMDLs** and that of **AK**. Note that the translation is not surjective: for every formula in the language of **TMDLs** there is a formula in the language of **AK**, but not vice versa. For example, there is no $\varphi \in \mathcal{L}$ such that $\mathcal{T}(\varphi)$ is of the form $\Box(\psi \to \mathcal{Q}\theta)$. Such formulas can be used to express 'deontic sufficiency': $\Box(\psi \to \mathcal{Q}\theta)$ means that the truth of ψ is sufficient for θ being a good person (see [21]).

4.2 Proof of the translation theorem for TMK

The proof of the translation theorem for **TMK**, Theorem 4.9, mostly follows that for the propositional case as laid out by Åqvist [3,4]. However, there are some important complications associated with the step to the term-modal level. The proof starts with Lemma 4.2, which shows that for every theorem of **TMK**, its translation is a theorem of **AK**. For the other direction it remains to prove that for every $\varphi \in \mathcal{L}$, if $\mathcal{T}(\varphi)$ is a theorem of **AK**, then φ is a theorem of **TMK** (Lemma 4.8). To prove this, we first define a model M^+ for every **TMK**-model M (Definition 4.3). This definition is significantly more complicated than its propositional counterpart, and as a result the following lemmas are as well. In the next step we prove what Åqvist calls the "easy lemma": each model constructed according to Definition 4.3 is an **AK**-model (Lemma 4.4). Then we prove Lemma 4.5, the lemma "on relations", stating that the accessibility relations in the **TMK**-model M correspond to that in the **AK**-model M^+. After the small auxiliary Lemma 4.6, we can prove the "crucial" Lemma 4.7: for any world w in the original **TMK** model, this world models φ iff the corresponding worlds in the **AK**-model M^+ model $\mathcal{T}(\varphi)$.

Lemma 4.2 *If* $\vdash_{\mathbf{TMK}} \varphi$*, then* $\vdash_{\mathbf{AK}} \mathcal{T}(\varphi)$.

Proof. The proof is a simple induction on the length of the derivation. For the base cases (i.e. derivations of length one), φ is an instance of one of the axiom schemes of **TMK**. We leave it to the reader to see that for each axiom scheme ψ of **TMK**, $\mathcal{T}(\psi)$ is a theorem of **AK** (see also [7]). For the induction step, it suffices to prove that if φ has been obtained by an application of one of the rules of **TMK** to some **TMK**-theorem(s), then $\mathcal{T}(\psi)$ is a theorem of **AK**. This is also safely left to the reader. □

Definition 4.3 Let $M = \langle W, \mathcal{A}, R, R^D, I\rangle$ be a **TMK**-model. We define $M^+ = \langle W^+, \mathcal{A}, R^Q, f, f^2, I^+\rangle$ as follows:
1. $W^+ = (W \times \mathcal{A}) \cup (W \times \mathcal{A} \times \mathcal{A})$
2. For all $w \in W$ and $p_1, p_2 \in \mathcal{A}$, $R^Q(\langle w, p_1\rangle) = R^Q(\langle w, p_1, p_2\rangle) = \{\langle w', p'\rangle \in W^+ \mid w' \in R(w, p')\} \cup \{\langle w'', p'', p'''\rangle \in W^+ \mid w'' \in R^D(w, p'', p''')\}$
3.1. For all $p \in \mathcal{A}$, $f(p) = \{\langle w, p\rangle \in W^+ \mid$ there exists a $w' \in W$ such that $w \in R(w', p)\}$
3.2. For all $p_1, p_2 \in \mathcal{A}$, $f^2(p_1, p_2) = \{\langle w, p_1, p_2\rangle \mid$ there exists a $w' \in W$ such that $w \in R^D(w', p_1, p_2)\}$
4.1. For all $\theta \in T$, $I^+(\theta) = I(\theta)$
4.2. For all $P \in \mathcal{P}$ and $\langle w, p\rangle, \langle w, p', p''\rangle \in W^+$, $I^+(P, \langle w, p\rangle) = I^+(P, \langle w, p', p''\rangle) = I(P, w)$

M^+ is defined such that for any **TMK**-model $M = \langle W, \mathcal{A}, R, R^D, I\rangle$, $w \in W$, $p \in \mathcal{A}$ and $\varphi \in \mathcal{L}$: $M, w \models \varphi$ iff $M^+, \langle w, p\rangle \models \mathcal{T}(\varphi)$ iff $M^+, \langle w, p, p'\rangle \models \mathcal{T}(\varphi)$ (Lemma 4.7). In what follows we will prove this formally, but now we first sketch an example to illustrate Definition 4.3.

Consider a **TMK**-model $M_1 = \langle W, \mathcal{A}, R, R^D, I\rangle$ such that:
$W = \{w_1, w_2\}$,
$\mathcal{A} = \{p_1, p_2\}$,
$R = \{\langle w_1, p_1, w_2\rangle, \langle w_1, p_1, w_1\rangle\}$ and
$R^D = \{\langle w_1, p_2, p_1, w_2\rangle\}$.
This model is graphically represented in Figure 1.

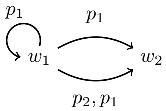


Fig. 1. The model M_1

$M_1^+ = \langle W^+, \mathcal{A}, R^Q, f, f^2, I^+\rangle$ is graphically represented in Figure 2. By clause 1 of Definition 4.3, W^+ contains the eight elements represented in Figure 2. By clause 2, for any $\langle w, p\rangle \in W^+$: $R^Q(\langle w, p\rangle)$ contains the worlds $\langle w', p'\rangle$ such that $w' \in R(w, p')$ and the worlds $\langle w', p', p''\rangle$ such that $w' \in R^D(w, p', p'')$. Similarly, for any $\langle w, p, p'\rangle \in W^+$: $R^Q(\langle w, p, p'\rangle)$ contains the worlds $\langle w', p''\rangle$ such that $w' \in R(w, p'')$ and the worlds $\langle w', p'', p'''\rangle$ such that $w' \in R^D(w, p'', p''')$.

So for every world that is a copy of w_1 (i.e. every world on the left side of Figure 2) there are three R^Q-accessible worlds: $\langle w_1, p_1\rangle$ (since $\langle w_1, p_1, w_1\rangle \in R$), $\langle w_2, p_1\rangle$ (since $\langle w_1, p_1, w_2\rangle \in R$) and $\langle w_2, p_2, p_1\rangle$ (since $\langle w_1, p_2, p_1, w_2\rangle \in R^D$). For every copy of w_2 there are no R^Q-accessible worlds, since no worlds are accessible from w_2 in model M_1.

By clause 3.1. of Definition 4.3, $\langle w, p\rangle \in f(p')$ iff $p = p'$ and there is a $w' \in W$ such that $w \in R(w', p)$. So $f(p_1) = \{\langle w_1, p_1\rangle, \langle w_2, p_1\rangle\}$, and $f(p_2) = \emptyset$. In Figure 2 we have designated the worlds in $f(p_1)$ by underlining the names of these worlds. By clause 3.2. of Definition 4.3, $f(p_2, p_1) = \langle w_2, p_2, p_1\rangle$. This is shown in Figure 2 by the double underlining of the name of this world. Note that $f(p_2) = f^2(p_1, p_2) = f^2(p_1, p_1) = f^2(p_2, p_2) = \emptyset$.

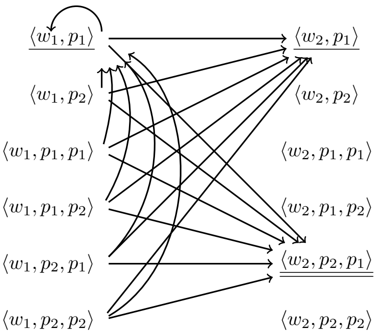

$\underline{\langle w_1, p_1\rangle}$
$\underline{\langle w_2, p_1\rangle}$
$\langle w_1, p_2\rangle$
$\langle w_2, p_2\rangle$
$\langle w_1, p_1, p_1\rangle$
$\langle w_2, p_1, p_1\rangle$
$\langle w_1, p_1, p_2\rangle$
$\langle w_2, p_1, p_2\rangle$
$\langle w_1, p_2, p_1\rangle$
$\underline{\underline{\langle w_2, p_2, p_1\rangle}}$
$\langle w_1, p_2, p_2\rangle$
$\langle w_2, p_2, p_2\rangle$

Fig. 2. The model M_1^+

To illustrate Lemma 4.7, take again model M_1 and suppose that $I(a) = p_1$, $I(b) = p_2$, $I(R, w_1) = \emptyset$, and $I(R, w_2) = p_2$. Then by the semantic clauses of **TMK**, $M_1, w_2 \models Rb$ and $M_1, w_1 \models \neg Rb \wedge \mathsf{O}_b^a Rb \wedge \neg \mathsf{O}_a Rb$. By Definition 4.1, $\mathcal{T}(Rb) = Rb$ and $\mathcal{T}(\neg Rb \wedge \mathsf{O}_b^a Rb \wedge \neg \mathsf{O}_a Rb) = \neg Rb \wedge \Box(\mathcal{Q}^2 ba \to Rb) \wedge \neg\Box(\mathcal{Q}a \to Rb)$. By clause 4.2. of Definition 4.3 and the semantic clauses of **AK**, $M_1^+, \langle w_1, p_1\rangle \models \Box(\mathcal{Q}^2 ba \to Rb)$ (since $f^2(I(b), I(a)) = \langle w_2, p_2, p_1\rangle$). Similarly, $M_1^+, \langle w_1, p_1\rangle \models \neg Rb \wedge \neg\Box(\mathcal{Q}a \to Rb)$. A similar line of reasoning can be made for every copy of w_1.

To prevent some possible confusion, we make two further clarificatory comments on clause 2. of Definition 4.3. First, note that to the left of the last identity symbol there is mention of p_1 and p_2, while on the right of that symbol there are p', p'', and p'''. In contrast, w occurs on both sides of the identity symbol. This is done intentionally. In each instance w refers to the same world, while p_1 and p_2 on the one hand and $p', p''p'''$ on the other are variables for different instances of quantification.

Second, note that for the proof it is important that each copy of a world has the same accessible worlds. We could, for example, not have written clause 2. as the following two clauses: For all $w \in W$ and $p_1, p_2 \in \mathcal{A}$, $R^Q(\langle w, p_1\rangle) = \{\langle w', p'\rangle \in W^+ \mid w' \in R(w, p')\}$ and $R^Q(\langle w, p_1, p_2\rangle) = \{\langle w'', p'', p'''\rangle \in W^+ \mid w'' \in R^D(w, p'', p''')\}$. If we had done so, then for example $M_1^+, \langle w_1, p_1, p_1\rangle \models \Box(\mathcal{Q}a \to)$ would have been the case, since there would have been no R^Q-accessible worlds that make $\mathcal{Q}a$ true. Given these clarifications, we now return to the proof.

Lemma 4.4 (Easy Lemma) *Given a* **TMK**-*model* $M = \langle W, \mathcal{A}, R, R^D, I\rangle$, $M^+ = \langle W^+, \mathcal{A}, R^Q, f, f^2, I^+\rangle$ *is an* **AK**-*model.*

Proof. It suffices to prove that M^+ satisfies all the conditions in Definition 3.1. In view of [7], this is safely left to the reader. □

Lemma 4.5 (Lemma on Relations) *Let* $M = \langle W, \mathcal{A}, R, R^D, I\rangle$ *and* $M^+ = \langle W^+, \mathcal{A}, R^Q, f, f^2, I^+\rangle$. *For all* $w, w' \in W$ *and* $p, p' \in \mathcal{A}$*:*
1. $w' \in R(w, p)$ *iff* $\langle w', p\rangle \in R^Q(\langle w, p\rangle)$ *and* $\langle w', p\rangle \in f(p)$
2. $w' \in R^D(w, p, p')$ *iff* $\langle w', p, p'\rangle \in R^Q(\langle w, p\rangle)$ *and* $\langle w', p, p'\rangle \in f^2(p, p')$

Proof. 1. and 2. both follow from clauses 2., 3.1. and 3.2. of Definition 4.3. □

Lemma 4.6 *Let* $M = \langle W, \mathcal{A}, R, R^D, I\rangle$ *and* $M^+ = \langle W^+, \mathcal{A}, R^Q, f, f^2, I^+\rangle$*, as in Definition 4.3. Then for all* $w \in W$ *and* $p_1, p_2, p_3, p_4 \in \mathcal{A}$*:* $R^Q(\langle w, p_1\rangle) = R^Q(\langle w, p_2\rangle) = R^Q(\langle w, p_3, p_4\rangle)$

Proof. This follows from clause 2. of Definition 4.3. □

Lemma 4.7 (Crucial Lemma) *Let* $M = \langle W, \mathcal{A}, R, R^D, I\rangle$ *be a* **TMK**-*model,* $w \in W$, $p, p_1 \in \mathcal{A}$ *and* $\varphi \in \mathcal{L}$*, then:*

$M, w \models \varphi$ *iff* $M^+, \langle w, p\rangle \models \mathcal{T}(\varphi)$ *iff* $M^+, \langle w, p, p_1\rangle \models \mathcal{T}(\varphi)$

Proof. The proof proceeds by induction on the complexity of φ. In view of Definitions 4.3 and 4.1, all cases can safely be left to the reader except for the cases where φ is of the form $\mathsf{O}_\theta\psi$ or $\mathsf{O}_\theta^\kappa\psi$.

We start with the case where φ is of the form $\mathsf{O}_\theta\psi$. We first prove that $M, w \models \mathsf{O}_\theta\psi$ iff $M^+, \langle w, p\rangle \models \mathcal{T}(\mathsf{O}_\theta\psi)$. For the left to right direction, assume that $M, w \models \mathsf{O}_\theta\psi$. By clauses 2. and 3.1. of Definition 4.3, $R^Q(\langle w, I(\theta)\rangle) \cap f(I(\theta)) \subseteq W \times \mathcal{A}$. Suppose that $\langle w', p'\rangle \in R^Q(\langle w, I(\theta)\rangle) \cap f(I(\theta))$. By clause 3.1. of Definition 4.3, $p' = I(\theta)$. By Lemma 4.5, $w' \in R(w, I(\theta))$. Since $M, w \models \mathsf{O}_\theta\psi$, $M, w' \models \psi$. By the Induction Hypothesis, $M^+, \langle w', I(\theta)\rangle \models \mathcal{T}(\psi)$. Thus, for all $\langle w', p'\rangle \in R^Q(\langle w, I(\theta)\rangle)$: if $\langle w', p'\rangle \in f(I(\theta))$, then $M^+, \langle w', I(\theta)\rangle \models \mathcal{T}(\psi)$. By Lemma 4.6, $R^Q(\langle w, I(\theta)\rangle) = R^Q(\langle w, p\rangle)$. Hence, by the semantic clauses of **AK**, $M^+, \langle w, p\rangle \models \Box(\mathcal{Q}\theta \to \mathcal{T}(\psi))$.

For the right to left direction, assume that $M^+, \langle w, p\rangle \models \Box(\mathcal{Q}\theta \to \mathcal{T}(\psi))$. Suppose that $w' \in R(w, I(\theta))$. By Lemma 4.5, $\langle w', I(\theta)\rangle \in R^Q(\langle w, I(\theta)\rangle) \cap f(I(\theta))$. By Lemma 4.6, $\langle w', I(\theta)\rangle \in R^Q(\langle w, p\rangle) \cap f(I(\theta))$. By the assumption,

$M^+, \langle w', I(\theta)\rangle \models \mathcal{T}(\psi)$. By the induction hypothesis, $M, w' \models \psi$. Thus, by the semantic clauses, $M, w \models \mathsf{O}_\theta\psi$.

For this case all that remains is to prove that $M^+, \langle w, p\rangle \models \mathcal{T}(\mathsf{O}_\theta\psi)$ iff $M^+, \langle w, p, p_1\rangle \models \mathcal{T}(\mathsf{O}_\theta\psi)$. By the definition of the translation, we have to prove that $M^+, \langle w, p\rangle \models \Box(\mathcal{Q}\theta \to \mathcal{T}(\psi))$ iff $M^+, \langle w, p, p_1\rangle \models \Box(\mathcal{Q}\theta \to \mathcal{T}(\psi))$. This follows immediately from the induction hypothesis, Lemma 4.6 and the semantic clauses.

For the case where φ is of the form $\mathsf{O}^\kappa_\theta\psi$, we first prove that $M, w \models \mathsf{O}^\kappa_\theta\psi$ iff $M^+, \langle w, p\rangle \models \mathcal{T}(\mathsf{O}^\kappa_\theta\psi)$.[10] For the left to right direction, assume that $M, w \models \mathsf{O}^\kappa_\theta\psi$. Suppose that $\langle w', p', p''\rangle \in R^Q(\langle w, I(\theta)\rangle) \cap f^2(I(\theta), I(\kappa))$. By clause 3.2. of Definition 4.3, $p' = I(\theta)$ and $p'' = I(\kappa)$. By Lemma's 4.5 and 4.6, $w' \in R^D(w, I(\theta), I(\kappa))$. Since $M, w \models \mathsf{O}^\kappa_\theta\psi$, $M, w' \models \psi$. By the induction hypothesis, $M^+, \langle w', I(\theta)\rangle \models \mathcal{T}(\psi)$. Thus, for all $\langle w', p', p''\rangle \in R^Q(\langle w, I(\theta)\rangle)$: if $\langle w', p', p''\rangle \in f^2(I(\theta), I(\kappa))$, then $M^+, \langle w', I(\theta)\rangle \models \mathcal{T}(\psi)$. By Lemma 4.6 and clause 3.2. of Definition 4.3, for all $\langle w', p'\rangle \in R^Q(\langle w, p\rangle)$: if $\langle w', p'\rangle \in f^2(I(\theta), I(\kappa))$, then $M^+, \langle w', p'\rangle \models \mathcal{T}(\psi)$ By the semantic clauses and the fact that $f^2(I(\theta), I(\kappa)) \subseteq W \times \mathcal{A} \times \mathcal{A}$, $M^+, \langle w, p\rangle \models \Box(\mathcal{Q}^2\theta, \kappa \to \mathcal{T}(\psi))$.

For the right to left direction, assume that $M^+, \langle w, p\rangle \models \Box(\mathcal{Q}^2\theta\kappa \to \mathcal{T}(\psi))$. Suppose that $w' \in R^D(w', I(\theta), I(\kappa))$. By Lemma 4.5, $\langle w', I(\theta), I(\kappa)\rangle \in R^Q(\langle w, I(\theta), I(\kappa)\rangle) \cap f^2(I(\theta), I(\kappa))$. By Lemma 4.6, $\langle w', I(\theta), I(\kappa)\rangle \in R^Q(\langle w, p\rangle) \cap f^2(I(\theta), I(\kappa))$. By the assumption, $M^+, \langle w', I(\theta), I(\kappa)\rangle \models \mathcal{T}(\psi)$. By the induction hypothesis, $M, w' \models \psi$. By the semantic clauses, $M, w \models \mathsf{O}^\kappa_\theta\psi$.

What remains to be proven for this case is that $M^+, \langle w, p\rangle \models \mathcal{T}(\mathsf{O}^\kappa_\theta\psi)$ iff $M^+, \langle w, p, p_1\rangle \models \mathcal{T}(\mathsf{O}^\kappa_\theta\psi)$. By the translation, this means that we have to prove that $M^+, \langle w, p\rangle \models \Box(\mathcal{Q}^2\theta\kappa \to \psi)$ iff $M^+, \langle w, p, p_1\rangle \models \Box(\mathcal{Q}^2\theta\kappa \to \psi)$. This follows immediately from Lemma 4.6 and the semantic clauses. □

Lemma 4.8 *For all $\varphi \in \mathcal{L}$, if $\vdash_{\mathbf{AK}} \mathcal{T}(\varphi)$, then $\vdash_{\mathbf{TMK}} \varphi$.*

Proof. Suppose that $\nvdash_{\mathbf{TMK}} \varphi$. By completeness, there is a **TMK**-model $M = \langle W, \mathcal{A}, R, R^D, I\rangle$ and a $w \in W$, such that $M, w \not\models \varphi$. By Lemmas 4.4 and 4.7, there is an **AK**-model $M^+ = \langle W^+, \mathcal{A}, R^Q, f, f^2, I^+\rangle$ and $\langle w, p\rangle \in W^+$, such that $M^+, \langle w, p\rangle \not\models \mathcal{T}(\varphi)$. By the soundness of **AK**, $\nvdash_{\mathbf{AK}} \mathcal{T}(\varphi)$. □

Theorem 4.9 $\{\varphi \in \mathcal{L} \mid\ \vdash_{\mathbf{TMK}} \varphi\} = \{\varphi \in \mathcal{L} \mid\ \vdash_{\mathbf{AK}} \mathcal{T}(\varphi)\}$

Proof. This follows immediately from lemmas 4.2 and 4.8. □

4.3 The translation theorems for the extensions

In this section we prove the translation theorem for each of the extensions of **TMK** that were defined in Section 2.4. The first column of Table 8 gives a list of axioms. Adding any combination of these axioms and the corresponding frame conditions (see Table 4) to **TMK** gives an extension of **TMK** (though note that

[10] From here on the proof differs significantly from that in [7], since we do not assume (DIU) for **TMK**.

we demand that either both or neither of (+) and (D+) are added[11]. To find the logic of which such an extension of **TMK** is the deontic fragment, simply add the axioms that are on the same lines in the second column of Table 8 (and their corresponding frame conditions) to **AK**. For example, **TMK** extended with (DIU), (+) and (D+) is the deontic fragment of **TMK** extended with (QQ) and (AT). In what follows we sketch how the proof of the translation theorem needs to be adapted for each of these extensions.

TM-axioms	**AK-axioms**	**Modification to definition of M^+**
(D)	(AQ)	no changes
(DD)	(ADQ)	no changes
(+) and (D+)	(AT)	add $\cup\{\langle w, p'\rangle \mid \langle w, p'\rangle \in W\}\cup$ $\{\langle w, p', p''\rangle \mid \langle w, p', p''\rangle \in W\}$ to 2.
(DIU)	(QQ)	add $\cup f(p_1)$ to 3.2.

Table 8
Term-modal and Andersonian-Kangerian axioms

Theorem 4.10 *For any combination of lines from Table 8, let* **TML** *be the extension of* **TMK** *obtained by adding the axioms (and corresponding frame conditions) from the first column to* **TMK***, and let* **AKL** *be the extension of* **AK** *obtained by adding the axioms (and corresponding frame conditions) from the second column to* **AK***. Then* $\{\varphi \in \mathcal{L} \mid \vdash_{\mathbf{TML}} \varphi\} = \{\varphi \in \mathcal{L} \mid \vdash_{\mathbf{AKL}} \mathcal{T}(\varphi)\}$.

Proof. For each of the logics, the proof is mostly analogous to that for **TMK** and **AK** in Section 4.2. We go through each of the lemmas and definitions in that section and describe where the proof changes substantially.

Proving the analogue of Lemma 4.2 is straightforward. The proof is analogous to that for Lemma 4.2, except that there are more cases to prove. For example, if the line with (D) and (AQ) is included, then we need to prove that $\vdash_{\mathbf{AKL}} \Box(\mathcal{Q}\theta \to \psi) \to \neg\Box(\mathcal{Q}\theta \to \neg\psi)$. This follows from (AQ) and the properties of normal modal logic. The other cases are analogous; each time the theorem follows from the axiom added to **AKL** and the properties of normal modal logic.

The Definition of M^+ needs to be adapted in some cases. These changes are described in the third column of Table 8. The easy Lemma (Lemma 4.4) is then safely left to the reader.

For the Lemma on relations (Lemma 4.5), only the extensions containing (+) and (D+) require further comment. For both 1. and 2. the left to right cases still follow immediately from clauses 2. and 3.1. or 3.2. For the right to left case of 1., assume $\langle w', p\rangle \in R^Q(\langle w, p\rangle)$ and $\langle w', p\rangle \in f(p)$. By (the modified) clause 2., either (A) $w' \in R(w, p)$ or (B) $w' = w$. In case of (B), by $\langle w, p\rangle \in f(p)$ and clause 3.1., there exists a $w'' \in W$ such that $w \in R(w'', p)$. By the frame

[11] This is not a problem, since any argument for the acceptance of (+) is in all likelihood also an argument for the acceptance of (D+).

condition corresponding to (+), $R(w'', p) = R(w, p)$. Hence $w \in R(w, p)$. For 2. the proof is analogous.

For the crucial Lemma (Lemma 4.7) there are no significant changes necessary for (D) or (DD). For (+) and (D+) there are also no significant changes, since it is still the case that $R^Q(\langle w, p\rangle) \subseteq W \times \mathcal{A}$ and $R^Q(\langle w, p, p'\rangle) \subseteq W \times \mathcal{A} \times \mathcal{A}$. For (DIU) the difference is significant, but the proof is entirely analogous to that for Lemma 4.16 in [7, pp. 94-95]. □

5 Conclusion

In this paper we have defined a family of **TMDLs** and a family of first-order logics inspired by the Andersonian-Kangerian view on deontic logic. We have gone on to show that the **TMDLs** are reducible to the Andersonian-Kangerian logics. In other words, we have shown that the **TMDLs** are a fragment (the deontic fragment) of the Andersonian-Kangerian logics. At first glance, the **TMDLs** may have seemed highly unorthodox because of the ternary/quaternary accessibility relations and the term-modal operators. However, the results in this paper show that **TMDLs** fit neatly with the canon of deontic logic.

On a conceptual level, it is often argued that the Andersonian-Kangerian reduction shows that the notion of obligation can be analysed in non-deontic terms. The same arguments can be made for the reduction presented in this paper. However, whether this argument is convincing depends on the reading of the constants, both in the propositional and the predicative case. If $\mathcal{Q}\theta$ is for example read as 'θ fulfils all of their obligations', then the reduction is merely an analysis of obligations in terms of obligations. Even if $\mathcal{Q}\theta$ is read as 'θ is a good person', then it might still be debated whether $\mathcal{Q}$ is not still a deontic constant (see [18] for a more detailed discussion).

There are different paths open for future work. In this paper we have used constant domains and have treated constants as rigid designators. It is still an open question whether this is actually the way to go for deontic logic, see for example [11,7]. If we can prove a similar reduction for variable domain semantics or non-rigid designators, then it might be argued that such questions about deontic logic can be reduced to the (much better studied) corresponding questions for alethic modal logic.

There are also a number of possible term-modal logics for which we have not yet proven the reduction to an Andersonian-Kangerian style system. Here one can think of the dyadic or non-normal **TMDLs** presented in [7]. Other examples might be found outside the domain of deontic logic itself. Term-modal versions of the 4, 5 and T-principles are relevant for doxastic and epistemic logic. The proofs for the reductions of such systems are not trivial. For example, for the 4-axiom it does not work to simply close R^Q in the definition of M^+ under transitivity. Things would be easier if we only had one accessibility relation in the term-modal models (which might be more applicable in an epistemic context anyway). Even more difficult would be the T-principle: if we close R^Q under reflexivity, then we lose Lemma 4.6.

Another, more philosophically inspired, possibility for future work is to explore the connection between **AKH** and virtue ethics. Very briefly, virtue ethics judges the correctness of an action not by the consequences of that action (as consequentialists might do), nor by whether that action corresponds to a previous set of rules (as a deontologist might). Instead, the virtue ethicist would say that an action is obligatory iff the action exemplifies some positive characteristic of the agent. Our definition of 'φ is obligatory for a' as 'φ is necessary for a being a good person' seems to fit this point of view rather well. Virtue ethics has mostly been overlooked in the literature on deontic logic, so it might be worthwhile to investigate whether this link can be substantiated. [12]

References

[1] Anderson, A. R., *A reduction of deontic logic to alethic modal logic*, Mind **67** (1958), pp. 100–103.

[2] Anderson, A. R., *Some nasty problems in the formal logic of ethics*, Noûs (1967), pp. 345–360.

[3] Åqvist, L., "Introduction to deontic logic and the theory of normative systems," Bibliopolis, 1987.

[4] Åqvist, L., *Deontic logic*, Handbook of Philosophical Logic: Volume 8 (2002), pp. 147–264.

[5] Barcan Marcus, R., *Iterated deontic modalities*, Mind **75** (1966), pp. 580–582.

[6] Fitting, M., L. Thalmann and A. Voronkov, *Term-modal logics*, Studia Logica **69** (2001), pp. 133–169.

[7] Frijters, S., "All doctors have an obligation to care for their patients: term-modal logics for ethical reasoning with quantified deontic statements," Ph.D. thesis, Ghent University (2021).

[8] Frijters, S. and T. De Coninck, *The manchester twins: Conflicts between directed obligations.*, in: *DEON*, 2021, pp. 166–182.

[9] Frijters, S., J. Meheus and F. Van De Putte, *Reasoning with rules and rights: term-modal deontic logic*, in: *New Developments in Legal Reasoning and Logic: From Ancient Law to Modern Legal Systems*, Springer, 2021 pp. 321–352.

[10] Frijters, S. and F. Van De Putte, *Classical term-modal logics*, Journal of Logic and Computation **31** (2021), pp. 1026–1054.

[11] Goble, L., *Quantified deontic logic with definite descriptions*, Logique et Analyse **37** (1994), pp. 239–253.

[12] Goble, L., *Prima facie norms, normative conflicts, and dilemmas*, Handbook of deontic logic and normative systems **1** (2013), pp. 241–351.

[13] Hilpinen, R., *Stig kanger on deontic logic*, Collected Papers of Stig Kanger with Essays on His Life and Work: Vol. II (2001), pp. 131–149.

[14] Hilpinen, R. and P. McNamara, *Deontic logic: A historical survey and introduction*, Handbook of deontic logic and normative systems **1** (2013), pp. 3–136.

[15] Kanger, S., *New foundations for ethical theory*, in: *Deontic Logic: Introductory and Systematic Readings*, Springer, 1971 pp. 36–58.

[16] Liberman, A. O., A. Achen and R. K. Rendsvig, *Dynamic term-modal logics for first-order epistemic planning*, Artificial Intelligence **286** (2020), p. 103305.

[17] Lindahl, L., *Stig kanger's theory of rights*, Collected Papers of Stig Kanger with Essays on His Life and Work: Vol. II (2001), pp. 151–171.

[12] I would like to thank one of the anonymous reviewers for this suggestion.

[18] McNamara, P. and F. Van De Putte, *Deontic Logic*, in: E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, 2022, Fall 2022 edition .

[19] Sawasaki, T., K. Sano and T. Yamada, *Term-sequence-modal logics*, in: *Logic, Rationality, and Interaction: 7th International Workshop, LORI 2019, Chongqing, China, October 18–21, 2019, Proceedings 7*, Springer, 2019, pp. 244–258.

[20] Thomason, R. H., *Some completeness results for modal predicate calculi*, Philosophical problems in logic: Some recent developments (1970), pp. 56–76.

[21] Van De Putte, F., *"That will do": logics of deontic necessity and sufficiency*, Erkenntnis **82** (2017), pp. 473–511.

Extensions and Variations of the DWE Framework with Applications

Paul McNamara [1]

Department of Philosophy, University of New Hampshire

Abstract

Building on McNamara's DWE (Doing Well Enough) framework, a substantial generalization of that framework extends its expressive resources by adding two ordering operators, changing the underlying semantic ordering's scope, and adding other operators making fuller use of the resulting expanded frames. A solution to DWE's "Disjunctive Supererogation Problem" is provided via two stronger semantic characterizations of permissible supererogation. The new framework can represent a divide in normative ethical theory between classical deontologists and consequentialists. It allows for modeling *supererogatory hole scenarios* and consequently, the *all or nothing problem scenarios.* We use the enriched resources to carefully explore modeling these problem cases, bringing greater resources to bear on them. It becomes clear that the classical deontological conflicts with consequentialism are close cousins to the just-mentioned problems with supererogation; each involves a growing divide in normative ethics over whether *something as good as something permissible must be permissible.* It is also made clear just how much impact it has if endorsed.

Keywords: DWE, supererogation, must, ought, the least one can do, good as ok is ok, supererogatory holes, all or nothing cases, deontology, consequentialism

Acknowledgement

I benefited from helpful discussions with Xavier Parent.

1 Introduction

In a variety of places (e.g. [12,13], [15]-[17]), I developed the DWE framework as a first approximation for modelling supererogation and associated concepts. However, there are artificial limitations in the semantic framework and substantial limitations in its expressive power, as well as in the logic of the supererogation operator itself. [2] We aim to rectify these lacks by expanding on and modifying the DWE framework, thereby generating a much richer and more

[1] paulm@unh.edu

[2] Some of these limitations have been previously noted in print (see especially [17], Section 5.10) and in prior conference presentations; in [17] and [18] expansions of framework are explored, but we do so here in a number of different ways making much fuller use of a generalization of the DWE semantic framework articulated there.

adequate framework. The resulting language is complex and we will focus on some semantic results to put them to use in modeling some classical debates between deontologists and optimizing consequentialists (for general background, see [1]), and in modeling two recent issues in ethical theory: Wessels' "supererogatory holes" (see especially [23,24]), and Horton's "all or nothing cases" [7] [3]. The framework can accommodate nuanced positions on these topics, as well as those that deny the basic intuitions often expressed (e.g. that there are supererogatory holes or that there can be all or nothing problem scenarios). We will see that the proposed operators for going beyond the call improve significantly on DWE's analysis, adding more nuance and subtlety while not succumbing to DWE's "Disjunctive Problem of Supererogation."

Section 2 sketches the classical DWE semantic framework and raises a number of questions about it to motivate our substantial expansions. Section 3.1 supplies those with a new language, which will include two ordering operators. [4] It then provides the semantics, resulting is a new framework, "DWE($\geq$)." Section 3.2 reflects on the expanded semantic framework that help us zero in on some payoffs: a resolution of the Disjunctive Supererogation Problem as well as modeling the aforementioned issues in ethical theory. In Section 4 we turn to a strengthened semantic framework, DWE$(\geq)^{\mathrm{G}}$, where we add a semantic constraint to the prior frames and then show its quite powerful reductive significance thereby highlighting some more payoffs of the general DWE($\geq$) framework. Section 5 concludes with some brief reflections on the framework and some future directions.

2 The DWE Framework and Some Limitations

2.1 Sketch of DWE's Language and Semantics [5]

Imagine a language for classic truth-functional propositional logic with the usual operators, $\neg$, &, $\vee$, $\rightarrow$, $\leftrightarrow$, and these primitive deontic operators of DWE added:

OBφ: It is *overridingly Obligatory* (for Jane) that φ. [6]

MIφ: The *Minimum* involves/implies (its being the case that) φ.

MAφ: The *Maximum* involves/implies (its being the case that) φ.

INφ: It is a matter of *Indifference* that φ. [7]

[3] Section 6.1 of [17] gives a brief critical exposition of Wessel's framework, and for just a couple of reactions to Horton's piece, see [19] and [20].

[4] [6] put an ordering relation to good use in discussing models for supererogation.

[5] See Section 5.6 of [17], as well as [12], [13], and [16] for expositions of this framework,

[6] We leave the intended relativization to an agent tacit in the remaining glosses on all operators. On the legitimacy of personal but non-agential readings of deontic operators, see especially Sections 1.3 of [14], briefly summarized in, Section 5.7 of [17], as well as the earlier [9]. See below on the qualifier "overridingly" in the reading here of **O**B.

[7] That is, the language envisioned is $\varphi ::= p \mid \neg\varphi \mid \varphi \rightarrow \varphi \mid \mathbf{O}\mathrm{B}\varphi \mid \mathbf{MI}\varphi \mid \mathbf{MA}\varphi \mid \mathbf{I}\mathrm{N}\varphi$, where $p \in S$ (sentence letters).

Here are a few of DWE's defined operators:

$\mathbf{P}\mathrm{E}\varphi \overset{def}{=} \neg\mathbf{O}\mathrm{B}\neg\varphi$ It is *Permissible* that φ.
$\mathbf{I}\mathrm{M}\varphi \overset{def}{=} \neg\mathbf{O}\mathrm{B}\neg\varphi$ It is *Impermissible* that φ.
$\mathbf{S}\mathrm{I}\varphi \overset{def}{=} \neg\mathbf{I}\mathrm{N}\varphi$ It is *Significant* that φ.
$\mathbf{B}\mathrm{C}\varphi \overset{def}{=} \mathbf{P}\mathrm{E}\varphi \;\;\&\;\; \mathbf{M}\mathrm{I}\neg\varphi$ It is *Beyond the Call* that φ.
$\mathbf{P}\mathrm{S}\varphi \overset{def}{=} \mathbf{P}\mathrm{E}\varphi \;\;\&\;\; \mathbf{M}\mathrm{A}\neg\varphi$ It is *Permissibly Suboptimal* that φ.

Next, let us define the frames for the DWE framework:

Definition 2.1 $F = \langle W, A, \succsim\rangle$, is a DWE Frame, where:

(i) $W \neq \phi$
(ii) $A \subseteq W^2$ and $\forall\ i\ \exists\ jAij$ (Seriality-A)
(iii) $\succsim\ \subseteq W^3$:
  a) $(k \succsim_i j$ or $j \succsim_i k)$ only if $(Aij\ \ \&\ \ Aik)$ [Confinement of $\succsim_i$ for A]
  b) $(Aij\ \ \&\ \ Aik)$ only if $(k \succsim_i j$ or $j \succsim_i k)$ [Connectivity of $\succsim_i$ for A]
  c) if $j \succsim_i k$ and $k \succsim_i l$ then $j \succsim_i l$ [Transitivity of $\succsim_i$]

Here is a pictorial representation:

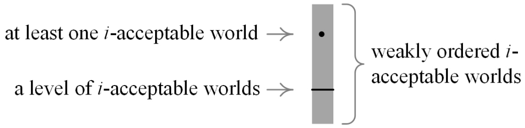

at least one i-acceptable world
a level of i-acceptable worlds
weakly ordered i-acceptable worlds

Fig. 1. DWE Frames

For simplicity, the frames are pictured as if there are upper and lower limits for the weak ordering. The context is classical: the focus is on modeling things that are *overridingly* obligatory, so on contexts where there is always an acceptable alternative (per world i)—one where all such overriding obligations are met. It is analytic that overriding obligations can't conflict with one another and overriding obligations provide the natural context for the new notions introduced in DWE since they reside primarily in the domain of what is optional. Note that for a world, i, the i-relative *ranking* relation is confined to the set of worlds that are i-acceptable. We define the following derivative notions (dropping parentheses where confusion is unlikely):

Definition 2.2 In a DWE Frame, $\langle W, A, \succsim\rangle$:

a) $A^i \overset{def}{=} \{j \in W : Aij\}$;

b) $j \approx_i k \overset{def}{=} j \succsim_i k \;\&\; k \succsim_i j$;

c) $j \succsim_i k \overset{def}{=} j \succsim_i k \;\&\; \neg(k \succsim_i j)$.

Models are defined as usual. Here are the truth conditions for the four primitive operators of DWE:

Definition 2.3 DWE's Basic Deontic Operator Truth-Conditions:
[**O**B] $M \models_i \mathbf{O}\mathrm{B}\varphi$: $\forall j(Aij \rightarrow M \models_j \varphi)$
[**MI**] $M \models_i \mathbf{MI}\varphi$: $\exists j(Aij \,\&\, \forall k(j \succsim_i k \rightarrow M \models_k \varphi))$
[**M**A] $M \models_i \mathbf{M}\mathrm{A}\varphi$: $\exists j(Aij \,\&\, \forall k(k \succsim_i j \rightarrow M \models_k \varphi))$
[**I**N] $M \models_i \mathbf{I}\mathrm{N}\varphi$: $\exists j[Aij \rightarrow \exists k(k \approx_i j \,\&\, M \models_k \varphi) \,\&$
$\exists k(k \approx_i j \,\&\, M \models_k \neg\varphi)]$

Here are truth conditions for our sample five defined operators:

Definition 2.4 Some of DWE's Derivative Truth-Conditions:
[**P**E] $M \models_i \mathbf{P}\mathrm{E}\varphi$: $\exists j(Aij \,\&\, M \models_j \varphi)$
[**I**M] $M \models_i \mathbf{I}\mathrm{M}\varphi$: $\exists j(Aij \rightarrow M \models_j \neg\varphi)$
[**S**I] $M \models_i \mathbf{S}\mathrm{I}\varphi$: $\exists j(Aij \,\&\, [\forall k(k \approx_i j \rightarrow M \models_k \varphi) \vee \forall k(k \succsim_i j \rightarrow$
$M \models_k \neg\varphi)])$
[**B**C] $M \models_i \mathbf{B}\mathrm{C}\varphi$: $\exists j(Aij \,\&\, M \models_j \varphi) \,\&\, \exists j[Aij \,\&\, \forall k(j \succsim_i k \rightarrow$
$M \models_k \neg\varphi)]$
[**P**S] $M \models_i \mathbf{P}\mathrm{S}\varphi$: $\exists j(Aij \,\&\, M \models_j \varphi) \,\&\, \forall j[Aij \,\&\, \exists k(k \approx_i j \,\&$
$M \models_k \neg\varphi)]$

Fig 2 Illustrates,

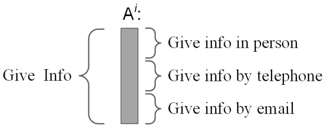


Fig. 2. DWE Illustration

Imagine you must provide a colleague with some delicate info. Suppose you can do so in exactly three ways, ordered on the right of the figure to reflect their value. All three delivery options on the right are *permissible* (they occur in acceptable worlds). You are *obligated* to provide the info (you do so in all acceptable worlds), what you ought to do (the *maximum*) involves giving the delicate info in person (you do so in the best acceptable worlds), the least you can do (the *minimum*) involves giving the info via email (you do so in the lowest ranked acceptable worlds); it is beyond the call to give the info in person or by phone since each is permissible but excluded from the minimal "acceptables" (acceptable worlds), and it is permissibly suboptimal to give the info by email or by phone, since these are precluded from the maximal acceptables. Each option is significant since there is a level of acceptable worlds (set of equi-ranked acceptables) throughout which that option occurs. Assume that it is a matter of indifference that you wear a belt, since all the acceptable levels of value can be achieved with or without wearing that.[8]

[8] We leave the logic determined by the semantics aside ([10]).

Let's pause to state DWE's *Disjunctive Supererogation Problem*:

Proposition 2.5 $\models$ *($\boldsymbol{I}M\psi$ & $\boldsymbol{B}C\varphi$) $\rightarrow$ $\boldsymbol{B}C(\varphi \vee \psi)$. [The Disjunctive Supererogation Problem]*

Proof. Assume $M \models_i \mathbf{I}\mathrm{M}\psi$ & $\mathbf{B}\mathrm{C}\varphi$, that is by def., $M \models_i \mathbf{I}\mathrm{M}\psi$ & $\mathbf{P}\mathrm{E}\varphi$ & $\mathbf{M}\mathrm{I}\neg\varphi$, so that (1) $\forall j(Aij \rightarrow \models_j \neg\psi)$, and (2) $\exists j(Aij$ & $M \models_j \varphi)$ and (3) $\exists j(Aij$ & $\forall k[j \gtrsim_i k \rightarrow M \models_k \neg\varphi])$. (2) entails (4) $\exists j(Aij$ & $M \models_j \varphi \vee \psi)$. Fixing j in (3), we get (5) Aij' and (6) $\forall k[j' \gtrsim_i k \rightarrow M \models_k \neg\varphi]$. For arbitrary k, suppose (7) $j' \gtrsim_i k$ & Aik. Then from (6), it follows that (8) $M \models_k \neg\varphi$. But (1) and (7) imply (9) $M \models_k \neg\psi$, and so we have (10) $M \models_k \neg(\varphi \vee \psi)$. But k was arbitrary in (7), so we have (11) $\forall k[j' \gtrsim_i k \rightarrow M \models_k \neg(\varphi \vee \psi)]$. But (11) conjoined to (5) yields (12) $\exists j(Aij$ & $\forall k[j \gtrsim_i k \rightarrow M \models_k \neg(\varphi \vee \psi)]$, and this conjoined to (4) is the truth-condition for $M \models_j \mathbf{B}\mathrm{C}(\varphi \vee \psi)$. □

To illustrate:

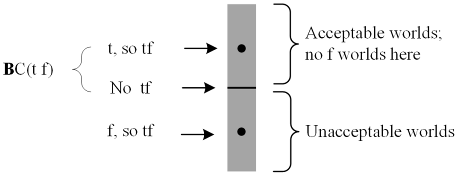


Fig. 3. The Disjunctive Supererogation Problem (for DWE)n

Tiny Tim is caught in a burning building and you pass by. The least you can do (e.g. go for help) precludes *rescuing Tiny Tim* (i.e. $\mathbf{M}\mathrm{I}\neg t$), which we assume is permissible ($\mathbf{P}\mathrm{E}t$), so per DWE, $\mathbf{B}\mathrm{C}t$. *Fanning the flames* is impermissible ($\mathbf{I}\mathrm{M}f$); so it is beyond the call for you to be such that either you rescue Tiny Tim or fan the flames ($\mathbf{B}\mathrm{C}(t \vee f)$). Upshot: Here going beyond the call seems to not really require that you go beyond the call. I have myself described the intuition behind my analysis of supererogation as "doing more good than you would have done had you done the least you could have done" (e.g. [17,16], but this condition seems to not be met by realizing $(t \vee f)$.

2.2 DWE Reinterpreted

Here we will reinterpret the DWE operators in anticipation of substantial revisions and expansions of that framework. Doing so will help pave the way for understanding various similar but distinct concepts and modeling them as distinct[9]. We begin by raising some questions about the DWE framework:

- Why not order the worlds generally per i and then integrate that ordering with a subset of i-acceptable worlds?

[9] This of course leaves open whether or not the concepts might be necessarily coextensive.

- Given the confinement of $\succsim_i$ to A^i, doesn't $\mathbf{I}$N amount to *indifference-among-permissible-alternatives*, not indifference per se (i.e. over a full ordering of acceptable and unacceptable worlds)?
- Is the permissible maximum the maximum in the full ordering? Is the best per se the acceptable best and what I ought to do?
- Is going beyond the call merely going beyond the permissible minimum or must it also be permissibly done?
- Given that there is a semantic ordering, would it be worthwhile to introduce a preference *operator* to reflect the semantic ordering?
- Can we solve DWE's *Disjunctive Supererogation Problem*?

Before exploring answers to these, let's add anticipatory qualifiers to DWE's naive operator labels (which we imagine as relativized to an agent).

DWE Operators with Anticipatory Qualifiers:

$\mathbf{O}\mathrm{B}\varphi$:	It is *overridingly Obligatory* that φ.
$\mathbf{M}\mathrm{I}^{\mathrm{P}}\varphi$:	The *Permissible Minimum* involves/implies φ.
$\mathbf{M}\mathrm{A}^{\mathrm{P}}\varphi$:	The *Permissible Maximum* involves/implies φ.
$\mathbf{I}\mathrm{N}^{\mathrm{P}}\varphi$:	It is *Indifferent* qua what is Permissible that φ.
$\mathbf{I}\mathrm{M}\varphi \overset{def}{=} \mathbf{O}\mathrm{B}\neg\varphi$	(It is *Impermissible* that φ.)
$\mathbf{P}\mathrm{E}\varphi \overset{def}{=} \neg\mathbf{O}\mathrm{B}\neg\varphi$	(It is *Permissible* that φ.)
$\mathbf{O}\mathrm{M}\varphi \overset{def}{=} \neg\mathbf{O}\mathrm{B}\varphi$	(It is *Omissible* that φ.)
$\mathbf{O}\mathrm{P}\varphi \overset{def}{=} \neg OB\varphi \;\&\; \neg\mathbf{O}\mathrm{B}\neg\varphi$	(It is *Optional* that φ.)
$\mathbf{N}\mathrm{O}\varphi \overset{def}{=} \mathbf{O}\mathrm{B}\varphi \vee \mathbf{O}\mathrm{B}\neg\varphi$	(It is *Non-Optional* that φ.)
$\mathbf{S}\mathrm{I}^{\mathrm{P}}\varphi \overset{def}{=} \mathbf{I}\mathrm{N}^{\mathrm{P}}\varphi$	(It is Significant *qua what is Permissible* that φ.)
$\mathbf{B}\mathrm{C}^{\mathrm{P}}\varphi \overset{def}{=} \mathbf{P}\mathrm{E}\varphi \;\&\; \mathbf{M}\mathrm{I}^{\mathrm{P}}\neg\varphi$	(It is *Permissibly Beyond the Call* (*the permissible minimum*) that φ.)
$\mathbf{P}\mathrm{S}^{\mathrm{P}}\varphi \overset{def}{=} \mathbf{P}\mathrm{E}\varphi \;\&\; \mathbf{M}\mathrm{A}^{\mathrm{P}}\neg\varphi$	(It is *Permissibly Suboptimal qua what is Permissible* that φ.)

The underlining introduces the intended qualified readings and the operators have a "$^{\mathrm{P}}$" indicator added.[10] This will allow us to more easily grasp conceptual distinctions to be made next.

3 DWE($\geq$)

3.1 The DWE($\geq$) Framework

We expand the DWE language with additional operators (but note that some DWE-basic operators will now be defined operators). Let's call the new framework "DWE($\geq$)".

[10] They reflect DWE's constraint of the ordering relation to A^i.

Definition 3.1 The Language of DWE(≥) follows: As with DWE, we assume a classical propositional logic language with these additional deontic operators:

The Primitive Operators of DWE(≥) and Their Readings:

$\varphi \geq \psi$:	It is *at least as good* that φ as that ψ.[11]
$\varphi \geq^{\mathrm{P}} \psi$:	It is *at least as good among the permissibles* that φ as that ψ.
$\mathbf{MI}^{\mathrm{P}}$:	The *Permissible Minimum* involves φ.
$\mathbf{MI}'\varphi$:	The *de facto Minimum (worst per se)* involves φ.
$\mathbf{B}\mathrm{M}^{\mathrm{P}*}\varphi$:	It is *unalterably permissibly Beyond the permissible Minimum* that φ[12].
$\mathbf{B}\mathrm{M}\varphi$:	It is *Beyond the permissible minimum* that φ.
$\mathbf{IN}^{\mathrm{P}}\varphi$:	It is *Indifferent among the permissibles* that φ.
$\mathbf{VI}\varphi$:	It is *Valuatively Indifferent* that φ[13].

Additional Defined Operators and Their Readings:

$\varphi > \psi \overset{def}{=} (\varphi \geq \psi)\ \&\ \neg(\psi \geq \varphi)$	It is *better that* φ than that ψ.
$\varphi \sim \psi \overset{def}{=} (\varphi \geq \psi)\ \&\ (\psi \geq \varphi)$	It is *just as good* that φ as that ψ.
$\varphi >^{\mathrm{P}} \psi \overset{def}{=} (\varphi \geq^{\mathrm{P}} \psi)\ \&\ \neg(\psi \geq^{\mathrm{P}} \varphi)$	It is *better among the Permissibles* that φ than that ψ.
$\varphi \sim^{\mathrm{P}} \psi \overset{def}{=} (\varphi \geq^{\mathrm{P}} \psi)\ \&\ (\psi \geq^{\mathrm{P}} \varphi)$	It is *just as good among the Permissibles* that φ as that ψ.
$\mathbf{B}\mathrm{M}^{\mathrm{P}} \overset{def}{=} \mathbf{P}\mathrm{E}\varphi\ \&\ \mathbf{B}\mathrm{M}\varphi$	It is *permissibly Beyond the Permissible Minimum* that φ.
$\mathbf{P}\mathrm{E}\varphi \overset{def}{=} (\varphi \geq^{\mathrm{P}} \varphi)$	It is *Permissible* that φ.
$\mathbf{I}\mathrm{M}\varphi \overset{def}{=} \neg(\varphi \geq^{\mathrm{P}} \varphi)$	It is *Impermissible* that φ.
$\mathbf{O}\mathrm{B}\varphi \overset{def}{=} \neg(\neg\varphi \geq^{\mathrm{P}} \neg\varphi)$	It is *overridingly Obligatory* that φ.
$\mathbf{O}\mathrm{M}\varphi \overset{def}{=} (\neg\varphi \geq^{\mathrm{P}} \neg\varphi)$	It is *Omissible* that φ.
$\mathbf{O}\mathrm{P}\varphi \overset{def}{=} (\varphi \geq^{\mathrm{P}} \varphi)\ \&\ (\neg\varphi \geq^{\mathrm{P}} \neg\varphi)$	It is *Optional* that φ.
$\mathbf{NO}\varphi \overset{def}{=} \neg(\neg\varphi \geq^{\mathrm{P}} \neg\varphi) \vee \neg(\varphi \geq^{\mathrm{P}} \varphi)$	It is *Non-Optional* that φ.
$\mathbf{MA}^{\mathrm{P}}\varphi \overset{def}{=} \varphi >^{\mathrm{P}} \neg\varphi$	The *Permissible Maximum* implies φ.
$\mathbf{MA}'\varphi \overset{def}{=} \varphi > \neg\varphi$	The *Maximum per se* implies φ.

[11] $\varphi \geq \psi$ on the proposed reading might seem to entail the truth of φ and ψ (see [3,4]), but here we wish to reflect the value the propositions would have from a world i if they were realized (i.e. "≥" has a subjunctive flavor). Similarly for $\geq^{\mathrm{P}}$. We will drop parentheses where the intent is clear for these dyadic operators.

[12] The "unalterably" qualifier indicates that the semantics will guarantee that any such φ can't be impermissibly realized, in contrast with for example $\mathbf{B}\mathrm{M}^{\mathrm{P}}\varphi$ (Proposition 3.5).

[13] That is, the language envisioned is $\varphi ::= p \mid \neg\varphi \mid \varphi \ \rightarrow\ \varphi \mid \varphi \geq \varphi \mid \varphi \geq^{\mathrm{P}} \varphi \mid \mathbf{MI}^{\mathrm{P}}\varphi \mid \mathrm{MI}'\varphi \mid \mathbf{B}\mathrm{M}^{\mathrm{P}*}\varphi \mid \mathbf{B}\mathrm{M}\varphi \mid \mathbf{IN}^{\mathrm{P}}\varphi \mid \mathbf{VI}\varphi$, where $p \in S$ (sentence letters).

$\mathbf{BC}^{\mathrm{P}}\varphi \overset{def}{=} \mathbf{PE}\varphi\ \&\ \ \mathbf{MI}^{\mathrm{P}}\neg\varphi$ It is *Beyond the call* (as in DWE) that φ.

$\mathbf{PS}^{\mathrm{P}}\varphi \overset{def}{=} \mathbf{PE}\varphi\ \&\ \mathbf{MA}^{\mathrm{P}}\neg\varphi$ It is *Suboptimal among the Permissibles* that φ.

$\mathbf{SO}\varphi \overset{def}{=} \mathbf{MA}'\neg\varphi$ It is *Sub-Optimal per se* that φ.

$\mathbf{SI}^{\mathrm{P}}\varphi \overset{def}{=} \mathbf{IN}^{\mathrm{P}}\varphi$ It is *Significant among the Permissibles* that φ.

$\mathbf{VS}\varphi \overset{def}{=} \neg\mathbf{VI}\varphi$ It is *Valuatively Significant* that φ.

$\mathbf{CI}\varphi \overset{def}{=} \mathbf{VI}\varphi\ \&\ \mathbf{IN}^{\mathrm{P}}\varphi$ It is *Completely Indifferent* that φ.

For DWE($\geq$), the only modifications of the DWE Frames are in clauses 3a and 3b, but we include all the clauses here:

Definition 3.2 $F = \langle W, A, \succsim\rangle$, is a DWE($\geq$) Frame, where:

1. $W \neq \phi$, [non-Emptiness]
2. $A \subseteq W^2$ and $\forall i \exists j Aij$, [Seriality-A]
3. $\succsim\ \subseteq W^3$:
   a) $j \succsim_i j$, [Reflexivity of $\succsim_i$][14]
   b) $k \succsim_i j$ or $j \succsim_i k$, [Connectivity of $\succsim_i$]
   c) $(j \succsim_i k\ \&\ k \succsim_i l) \rightarrow j \succsim_i l$ [Transitivity of $\succsim_i$]

Now, for each i in W, the connected weak ordering is *not confined* to A^i, but ranges over *all* of W. This is the key change, but as we shall see, the enrichment it brings is quite substantial: it opens a great deal of space for expressing different notions and positions of interest in ethical theory, and toward the end we will be looking at an additional fundamental constraint that has substantial additional impact. Fig 4 pictures some key potentials in the frames.

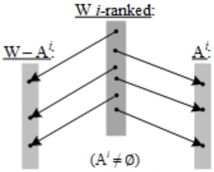


Fig. 4. DWE($\geq$) Frames

Here the central bar is the weak ordering of W per i, the right bar is the sub-ordering of the i-acceptables, the left bar is the sub-ordering of the i-unacceptables. As the arrows and angling suggest, the selection is *order-preserving*: for any two worlds j, k both appearing among one of the two sets

[14] Clause a' follows from Clause b' of course.

of non-central worlds, the ordering relation they have there just is the one they have in the i-ranking of W; but as indicated, there can be gaps.

The changes in the frames call for some qualifications in the truth conditions for the classical DWE operators to restore what was their intended for them, and we also must introduce new clauses for the new operators. All the familiar basic clauses for the traditional SDLish operators will hold (but that claim must be justified since these are all now defined operators and so their explicit truth conditions are those given via the conditions for their definiens). In the case of the non-SDL-ish operators of classical DWE that were interpreted via an ordering relation *confined* to A^i, those conditions must be qualified to restore the intent.

Assume we have sufficient clauses for the truth-functional operators: ¬, &, ∨, →, ↔; the truth conditions for DWE(≥)'s basic operators follow:

Basic DWE(≥) Deontic Operator Truth-Conditions:

[≥] $M \models_i \varphi \geq \psi$	$\forall j[M \models_j \psi \ \rightarrow \ \exists k(k \succsim_i j \ \& \ M \models_k \varphi)]$
[$\geq^{\mathrm{P}}$] $M \models_i \varphi \geq^{\mathrm{P}} \psi$	$\exists j(M \models_j \varphi \ \& \ Aij) \ \& \ \forall k[(M \models_k \psi \ \& \ Aik) \ \rightarrow \ \exists l(M \models_l \psi \ \& \ Ail \ \& \ l \succsim_i k)]$
[$\mathbf{MI}^{\mathrm{P}}$] $M \models_i \mathbf{MI}^{\mathrm{P}}\varphi$:	$\exists j(Aij \ \& \ \forall k[j \succsim_i k \ \& \ Aik) \ \rightarrow \ M \models_k \varphi])$
[$\mathbf{MI}'$] $M \models_i \mathbf{MI}'\varphi$:	$\exists j \forall k(j \succsim_i k \ \rightarrow \ M \models_k \varphi])$.[15]
[$\mathbf{B}\mathrm{M}^{\mathrm{P}^*}$] $M \models_i \mathbf{B}\mathrm{M}^{\mathrm{P}^*}\varphi$:	$\exists j(Aij \ \& \ M \models_j \varphi) \ \& \ \exists j'(Aij' \ \& \ \forall k[(j' \succsim_i k \ \& \ Aik) \ \rightarrow \ [M \models_k \neg\varphi \ \& \ \forall l(M \models_l \varphi \ \rightarrow \ (l \succ_i j' \ \& \ Ail))]])$
[$\mathbf{B}\mathrm{M}$] $M \models_i \mathbf{B}\mathrm{M}\varphi$:	$\exists j' M \models_{j'} \varphi \ \& \ \exists j[Aij \ \& \ \forall k[(j \succsim_i k \ \& \ Aik) \ \rightarrow \ [M \models_k \neg\varphi \ \& \ \forall l(M \models_i \varphi \ \rightarrow \ l \succ_i j)]])]$
[$\mathbf{I}\mathrm{N}^{\mathrm{P}}$] $M \models_i \mathbf{I}\mathrm{N}^{\mathrm{P}}\varphi$:	$\forall j[Aij \ \rightarrow \ \exists k(k \approx_i j \ \& \ Aik \ \& \ M \models_k \varphi) \ \& \ \exists k(k \approx_i j \ \& \ Aik \ \& \ M \models_k \neg\varphi)]$
[$\mathbf{VI}$] $M \models_i \mathbf{VI}\varphi$:	$\forall j[\exists k(k \approx_i j \ \& \ M \models_k \varphi) \ \& \ \exists k(k \approx_i j \ \& \ M \models_k \neg\varphi)]$

So $\models_i \varphi \geq \psi$ will hold (in a model M) iff "φ tracks ψ" in the $\succsim_i$ ordering, that is, for every ψ-world, there is a φ-world i-ranked (per $\succsim_i$) at least as high; whereas $\models_i \varphi \geq^{\mathrm{P}} \psi$ will hold iff "φ tracks ψ" among the i-acceptables (the i-acceptable worlds) per $\succsim_i$—more explicitly, there is an i-acceptable φ-world and for all i-acceptable ψ-worlds (if any), there is some i-acceptable φ-world i-ranked (at least) as high. $\models_i \mathbf{MI}^{\mathrm{P}}\varphi$ holds iff there is an i-acceptable world j such that all *i-acceptables* i-ranked as low are φ-worlds; whereas $M \models_i \mathbf{MI}'\varphi$ holds iff there is a world j such that *all worlds* i-ranked as low are φ-worlds.[16] $M \models_i \mathbf{B}\mathrm{M}^{\mathrm{P}^*}\varphi$ holds iff there is at least one i-acceptable φ-world and there is an i-acceptable world j such that a) from there on down among the i-acceptables φ is precluded and b) every φ-world is both i-ranked above j and i-acceptable.

[15] We cannot define $\mathbf{MI}\varphi$ as $\neg\varphi > \varphi$ for "$\neg\varphi > \varphi$" says $\neg\varphi$ at its best outranks any φ worlds. We could introduce a mirror image "≤" operator that says $\varphi \leq \psi$ holds iff for all ψ worlds, there is a φ-world ranked at least *as low*, but for our contrastive purposes we only need an **MI** operator.

[16] Since $j \succsim_i j$, in both cases, j must itself be a φ-world, and so only if, respectively, φ is permissible, φ is possible.

$M \models_i \mathbf{B}\mathrm{M}\varphi$ holds iff there is a φ-world (i.e. φ is possible) and there is an i-acceptable world j such that a) from there on down among the i-acceptables φ is precluded and b) every φ-world is i-ranked above j (whether i-acceptable or not). This allows for *impermissibly* exceeding the permissible minimum, as needed to map some conceptual space in current ethical theory. $\models_i \mathbf{I}\mathrm{N}^{\mathrm{P}}\varphi$ holds iff among the i-acceptable worlds, for every level (equivalence class per $\approx_i$) there is a φ-world and a $\neg\varphi$-world at that level; i.e. the realization of no i-acceptable level of value hinges on φ's status. Similarly for $\models_i \mathbf{V}\mathrm{I}\varphi$ except with no restriction to the i-acceptable worlds—for every i-level per se, there is a φ-world and a $\neg\varphi$-world at that level.

Here are truth conditions for the defined operators:[17]

Defined Deontic Operator Truth Conditions:

$[>]M \models_i \varphi > \psi$:	$\exists j[M \models_i \varphi \;\&\; \forall k(M \models_k \psi \rightarrow j \succ_i k)]$
$[\sim]M \models_i \varphi > \psi$:	$\exists j[M \models_i \psi \rightarrow \exists k(k \succsim_i j \;\&\; M \models_k \varphi)] \;\&$ $\forall j[M \models_j \varphi \rightarrow \exists k(k \succsim_i j \;\&\; M \models_k \psi)]$
$[>^{\mathrm{P}}]M \models_i \varphi >^{\mathrm{P}} \psi$:	$\exists j[M \models_j \varphi \;\&\; Aij \;\&\; \forall k((M \models_k \psi \;\&\; Aik \rightarrow j \succ_i k)]$
$[\sim^{\mathrm{P}}]M \models_i \varphi \sim^{\mathrm{P}} \psi$:	$\forall j[(M \models_j \psi \;\&\; Aij) \rightarrow \exists k(k \succsim_i j \;\&\; M \models_k \varphi \;\& Aij)] \;\&\; \forall j[(M \models_j \varphi \;\&\; Aij) \rightarrow \exists k(k \succsim_i j \;\&\; M \models_k \psi \;\&\; Aij$
$[\mathbf{P}\mathrm{M}]M \models_i \mathbf{P}\mathrm{E}\varphi$:	$\exists j(Aij \;\&\; M \models_j \varphi)$
$[\mathbf{I}\mathrm{M}]M \models_i \mathbf{I}\mathrm{M}\varphi$:	$\forall j(Aij \rightarrow M \models_j \neg\varphi)$
$[\mathbf{O}\mathrm{B}]M \models_i \mathbf{O}\mathrm{B}\varphi$:	$\forall j(Aij \rightarrow M \models_j \varphi)$
$[\mathbf{O}\mathrm{M}]M \models_i \mathbf{O}\mathrm{M}\varphi$:	$\exists j(Aij \;\&\; M \models_j \neg\varphi)$
$[\mathbf{O}\mathrm{P}]M \models_i \mathbf{O}\mathrm{P}\varphi$:	$\exists j(Aij \;\&\; M \models_j \varphi) \;\&\; \exists j(Aij \;\&\; M \models_j \neg\varphi)$
$[\mathbf{N}\mathrm{O}]M \models_i \mathbf{N}\mathrm{O}\varphi$:	$\exists j(Aij \rightarrow M \models_j \varphi) \vee \forall j(Aij \rightarrow M \models_j \neg\varphi)$
$[\mathbf{M}\mathrm{A}^{\mathrm{P}}]M \models_i \mathbf{M}\mathrm{A}^{\mathrm{P}}\varphi$:	$\exists j(Aij \;\&\; \forall k[k \succsim_i j \;\&\; Aik) \rightarrow M \models_k \varphi)$
$[\mathbf{M}\mathrm{A}']M \models_i \mathbf{M}\mathrm{A}'\varphi$:	$\exists j \forall k(k \succsim_i j \rightarrow M \models_k \varphi)$
$[\mathbf{B}\mathrm{C}^{\mathrm{P}}]M \models_i \mathbf{B}\mathrm{C}^{\mathrm{P}}\varphi$:	$\exists j(Aij \;\&\; M \models_j \varphi) \;\&\; \exists j(Aij \;\&\; \forall k[(j \succsim_i k \;\&\; Aik) \rightarrow M \models_k \neg\varphi])$
$[\mathbf{B}\mathrm{M}^{\mathrm{P}}]M \models_i \mathbf{B}\mathrm{M}^{\mathrm{P}}\varphi$:	$\exists j(Aij \;\&\; M \models_j \varphi) \;\&\; \exists j[Aij \;\&\; \forall k((j \succsim_i k \;\&\; Aik) \rightarrow [M \models_k \neg\varphi \;\&\; \forall l(M \models_l \varphi \rightarrow l \succ j)])]$
$[\mathbf{P}\mathrm{S}^{\mathrm{P}}]M \models_i \mathbf{P}\mathrm{S}^{\mathrm{P}}\varphi$:	$\exists j(Aij \;\&\; M \models_j \varphi) \;\&\; \exists j[Aij \;\&\; \forall k[(k \succsim_i j \;\&\; Aik) \rightarrow [M \models_k \neg\varphi])$
$[\mathbf{S}\mathrm{O}]M \models_i \mathbf{S}\mathrm{O}\varphi$:	$\exists j(M \models_j \varphi) \;\&\; \exists j \forall k(k \succsim_i j \rightarrow M \models_k \neg\varphi)$
$[\mathbf{S}\mathrm{I}^{\mathrm{P}}]M \models_i \mathbf{S}\mathrm{I}^{\mathrm{P}}\varphi$:	$\exists j(Aij \;\&\; [\forall k((k \approx_i j \;\&\; Aik) \rightarrow M \models_k \varphi) \vee \forall k((k \approx_i j \;\&\; Aik) \rightarrow [M \models_k \neg\varphi)])$
$[\mathbf{V}\mathrm{S}]M \models_i \mathbf{V}\mathrm{S}\varphi$:	$\exists j[\forall k(k \approx_i j \rightarrow M \models_k \varphi) \vee \forall k(k \approx_i j \rightarrow M \models_k \neg\varphi)]$
$[\mathbf{C}\mathrm{I}]M \models_i \mathbf{C}\mathrm{I}\varphi$:	$\forall j[Aij \rightarrow \exists k(k \approx_i j \;\&\; Aik \;\&\; M \models_k \varphi) \;\&$ $\exists k(k \approx_i j \;\&\; Aij \;\&\; M \models_k \neg\varphi)] \;\&\; \forall j[\exists k(k \approx_i j \;\&\; M \models_k \varphi) \;\&\; \exists k(k \approx_i j \;\&\; M \models_k \neg\varphi)]$

[17] Most of these are *derivative* truth-conditions, and so need justificatory proofs that the truth conditions via their definiens and the explicit conditions for those are equivalent to the ones given above. These are left out given space constraints.

$\models_i \varphi > \psi$ holds iff there is a φ-world i-ranked higher than any ψ-world, $\models_i \varphi \sim \psi$ holds iff φ tracks ψ and vice versa. $\models_i \varphi >^{\mathrm{P}} \psi$ holds iff there is an i-acceptable φ-world that outranks every i-acceptable ψ-world, and $\models_i \varphi \sim^{\mathrm{P}} \psi$ holds iff for every i-acceptable ψ-world there is an i-acceptable φ-world i-ranked as high, and vice versa.[18] For the next six defined SDL-style operators, standard truth conditions are given since they are easily derivable, so that for example $\models_i \mathbf{PE}\varphi$ holds iff $\models_i \varphi \geq^{\mathrm{P}} \varphi$ holds by definition, but that is equivalent to there is an i-acceptable φ-world. $\models_i \mathbf{M}\mathrm{A}^{\mathrm{P}}\varphi$ holds iff there is an i-acceptable world where *all i-acceptable worlds* i-ranked as high are φ-worlds; while $\models_i \mathrm{MA}'\varphi$ holds iff there is *a world* such that *all worlds* i-ranked as high are φ-worlds. $\models_i \mathbf{B}\mathrm{C}^{\mathrm{P}}\varphi$ holds iff there is an i-acceptable φ-world and there is an i-acceptable world j such that all i-acceptables i-ranked as low exclude φ. $\mathbf{B}\mathrm{M}^{\mathrm{P}}\varphi$ holds iff the conditions for $\mathbf{B}\mathrm{M}\varphi$ mentioned above hold as well as those for $\mathbf{P}\mathrm{E}\varphi$. $\models_i \mathbf{PS}^{\mathrm{P}}\varphi$ holds iff there is an i-acceptable φ-world and there is an i-acceptable world j such that all i-acceptables i-ranked as high exclude φ. In contrast $\models_i \mathbf{SO}\varphi$ holds iff there is a φ-world (φ is possible) and there is a world j such that all worlds i-ranked as high exclude φ. $\models_i \mathbf{SI}^{\mathrm{P}}\varphi$ holds iff $\models_i \neg\mathbf{IN}^{\mathrm{P}}\varphi$, that is, iff there is an i-acceptable world j such that either all i-acceptable worlds i-ranked equally with j are φ-worlds or all i-acceptable worlds i-ranked equally with j are $\neg\varphi$-worlds—the status of φ is essential to realizing some i-acceptable level; whereas $\models_i \mathbf{VS}\varphi$ holds iff there is a world j such that either all worlds i-ranked equally with j are φ-worlds or all worlds i-ranked equally with j are $\neg\varphi$ -worlds—the status of φ is essential to realizing some i-level of value (i-acceptable or not); Finally, $\models_i \mathbf{CI}\varphi$ holds iff the conditions above for both $\models_i \mathbf{VI}\varphi$ and $\models_i \mathbf{IN}^{\mathrm{P}}\varphi$ are met.

3.2 Some Applications and Reflections on the DWE($\geq$) Framework

Here we highlight some key features of the semantic framework, but with a special focus on some applications to ethical theory.

First, let's note that an option's being an unalterably permissible case of going beyond the permissible minimum entails that it is beyond the permissible minimum (whether permissible or not) and entails that it is permissible as well:

Proposition 3.3 $\models \boldsymbol{B}M^{P^*}\varphi \rightarrow (\boldsymbol{B}M\varphi \;\&\; \boldsymbol{B}M^{P}\varphi)$

Proof. Assume $M \models_i \mathbf{B}\mathrm{M}^{\mathrm{P}^*}\varphi$, that is (1) $\exists j(Aij \;\&\; M \models_j \varphi)$ & (2) $\exists j[(Aij \;\&\; \forall k[(j \succsim_i k \;\&\; Aik) \rightarrow [M \models_k \neg\varphi \;\&\; \forall l(M \models_l \varphi \rightarrow (l \succ j \;\&\; Ail))]]])$. Note that (2) at once implies the sightly weaker (2') $\exists j(Aij \;\&\; \forall k[(j \succsim_i k \;\&\; Aik) \rightarrow [M \models_k \neg\varphi \;\&\; \forall l[(M \models_l \varphi) \rightarrow l \succ j]])$ by predicate logic alone. But (1) also implies the weaker (1') $\exists j' M \models_j \varphi$,

[18] Although $\varphi \sim \psi$ and $\varphi \sim^{\mathrm{P}} \psi$ might be said to represent a sort of indifference, they are strictly weaker than $\mathbf{VI}\varphi$ and $\mathbf{I}\mathrm{N}\varphi$, respectively. Imagine all i-levels, and all i-acceptable levels, can be placed one-to-one with 1, 2, 3, ..., and for each odd-mapped such level there are all φ-worlds and for each even-mapped such level, there are all $\neg\varphi$-worlds; then $\varphi \sim \psi$ and $\varphi \sim^{\mathrm{P}} \psi$ are true, but $\mathbf{VI}\varphi$ and $\mathbf{IN}\varphi$ are false. That $\mathbf{VI}\varphi$ and $\mathbf{IN}\varphi$ are respectively as strong is easily seen.

which conjoined to (2') is the truth condition for $M \models_i \mathbf{B}\mathrm{M}\varphi$. Also, since (1) is the truth condition for $M \models_i \mathbf{P}\mathrm{E}\varphi$, by definition, we have $\models_i \mathbf{B}\mathrm{M}^{\mathrm{P}}\varphi$. □

Corollary 3.4 $\models \boldsymbol{B}M^{P^*}\varphi \rightarrow \boldsymbol{P}E\varphi$.

In Section 2.1 we explained DWE's *Disjunctive Supererogation Problem* and proved the following (now merely adapted to our alternative notation):

$\models (\mathbf{I}\mathrm{M}\psi \;\&\; \mathbf{B}\mathrm{C}^{\mathrm{P}}(\varphi \vee \psi)$. [*The Disjunctive Supererogation Problem*]

Why does DWE have this problem? Given $\models_i \mathbf{B}\mathrm{C}^{\mathrm{P}}\varphi$, nothing guarantees that φ's realization will exceed the permissible minimum, only that it can (i.e. there is *an acceptable* world where it does). In contrast, none of our three variants of going beyond the permissible minimum fall prey to this problem:

Proposition 3.5 $\not\models (\boldsymbol{I}M\psi \;\&\; \boldsymbol{B}M^{P^*}\varphi) \rightarrow \boldsymbol{B}M^{P^*}(\varphi \vee \psi)$. *[No Disjunctive Supererogation Problem]*

Proof. Let M be such that: $W = \langle i, j, k\rangle$, $A^i = \{i, j\}$, $M \models_i \varphi \;\&\; \neg\psi$, $M \models_j \neg\varphi \;\&\; \neg\psi$, $M \models_k \neg\varphi \;\&\; \psi$, and $i \succ_i j \succ_i k$ (with reflexivity of $\succsim_i$ for i, j, k)[19]. $M \models_i \mathbf{I}\mathrm{M}\psi$ is obvious. For $M \models_i \mathbf{B}\mathrm{M}^{\mathrm{P}^*}\varphi$: $Aii \;\&\; M \models_i \varphi$, so a) $\exists j(Aij \;\&\; M \models_j \varphi)$. Also Aij and so both $j \succsim_i j$ and Aij and $M \models_j \neg\varphi$ (but not $j \succsim_i i$ and not Aik); and $M \models_i \varphi$ but not so for j or k, and $i \succ_i j \;\&\; Aii$. So b) $\exists j'(Aij' \;\&\; \forall k[(j' \succsim_i k \;\&\; Aik) \rightarrow [M \models_k \neg\varphi \;\&\; \forall l(M \models_l \varphi \rightarrow (l \succ j' \;\&\; Ail))$, and b) conjoined to a) yields $M \models_i \mathbf{B}\mathrm{M}^{\mathrm{P}^*}\varphi$. For $M \not\models_i \mathbf{B}\mathrm{M}^{\mathrm{P}^*}(\varphi \vee \psi)$: First, i (and only i) does satisfy the condition on j in a') $\exists j(Aij \;\&\; M \models_j (\varphi \vee \psi))$. But each world fails to satisfy the condition of j' in b') $\exists j'(Aij' \;\&\; \forall k[(j' \succsim_i k \;\&\; Aik) \rightarrow [M \models_k \neg(\varphi \vee \psi) \;\&\; \forall l(M \models_l (\varphi \vee \psi) \rightarrow (l \succ j' \;\&\; Ail))]])$; for i does not meet it since although Aii and $i \succsim_i i$ hold, $M \models_i \neg(\varphi \vee \psi)$ fails; nor does k meet it since $\neg Aik$; nor does j meet it since although Aij, $j \succsim_i j$, $M \models_j \neg(\varphi \vee \psi)$, and $M \models_k (\varphi \vee \psi)$ hold, $k \succ j$ fails (as does Aik). □

Corollary 3.6 $\not\models (\boldsymbol{I}M\psi \;\&\; \boldsymbol{B}M\varphi) \rightarrow \boldsymbol{B}M(\varphi \vee \psi)$.[20]

Corollary 3.7 $\not\models (\boldsymbol{I}M\psi \;\&\; \boldsymbol{B}M^{P}\varphi) \rightarrow \boldsymbol{B}M^{P}(\varphi \vee \psi)$[21]

[19] For invalidating models, I will assume reflexivity without stating it is so, and likewise I will not remark on transitivity, connectivity, and seriality being satisfied when obvious.

[20] Given the proof of $\models \mathbf{B}\mathrm{M}^{\mathrm{P}^*}\varphi \rightarrow \mathbf{B}\mathrm{M}\varphi$ in Proposition 3.3, we need only show that $\models_i \mathbf{B}\mathrm{M}(\varphi \vee \psi)$ fails in the model for Proposition 3.5. Tracing through the model makes that apparent: in brief, the only world in the model that can fit the condition on j in $\exists j[Aij \;\&\; \forall k((j \succsim_i k \;\&\; Aik) \rightarrow M \models_k \neg(\varphi \vee \psi)$ is j itself, but j does not meet condition $\forall l(M \models_l (\varphi \vee \psi) \rightarrow l \succ j)$ since $\models_k (\varphi \vee \psi)$ yet $j \succ k$.

[21] Recall $\mathbf{B}\mathrm{M}^{\mathrm{P}}\varphi \overset{def}{=} \mathbf{P}\mathrm{E}\varphi \;\&\; \mathbf{B}\mathrm{M}\varphi$. Since $\models \mathbf{B}\mathrm{M}^{\mathrm{P}^*}\varphi \rightarrow \mathbf{B}\mathrm{M}\varphi$, we need only note that $\models_i \mathbf{P}\mathrm{E}\varphi$ holds in the model for Proposition 3.5 (obvious) and that $\models_i \mathbf{B}\mathrm{M}(\varphi \vee \psi)$ fails in the model for the reasons just noted for the first corollary. So $\models_i \mathbf{I}\mathrm{M}\psi \;\&\; \mathbf{B}\mathrm{M}^{\mathrm{P}}\varphi$ holds, but not so for $\models_i \mathbf{P}\mathrm{E}(\varphi \vee \psi) \;\&\; \mathrm{BM}(\varphi \vee \psi)$.

Corollary 3.8 $\not\models \boldsymbol{B}C^P\varphi \ \rightarrow \boldsymbol{B}M\varphi.$[22]

Observation 3.1 *Neither* $\boldsymbol{B}M$*,* $\boldsymbol{B}M^P$*, nor* $\boldsymbol{B}M^{P^*}$ *fall prey to the disjunctive supererogation problem.*

$\mathbf{B}\mathrm{M}\varphi$, $\mathbf{B}\mathrm{M}^{\mathrm{P}}\varphi$ and $\mathbf{B}\mathrm{M}^{\mathrm{P}^*}\varphi$ each require that to be beyond the permissible minimum at all, φ must guarantee the realization of more good than the permissible minimum does, with $\mathbf{B}\mathrm{M}^{\mathrm{P}}\varphi$ entailing that φ *can* be permissibly realized, and with $\mathrm{BM}^{\mathrm{P}^*}\varphi$ guaranteeing that φ *can only* be permissibly realized. All three operators fit my frequent gloss on going beyond the call as "doing more good than the permissible minimum", although I assumed it must be permissible to do so as well, which $\mathbf{B}\mathrm{M}^{\mathrm{P}}$ and $\mathbf{B}\mathrm{M}^{\mathrm{P}^*}$ entail, unlike $\mathbf{B}\mathrm{M}\varphi$. Let's we verify the latter claim—that we can model *impermissibly* going beyond the permissible minimum—next.

Proposition 3.9 $\not\models \boldsymbol{B}M\varphi \ \rightarrow \boldsymbol{P}E\varphi$

Proof. Let $W = \{i, j\}$, $A^i = \{j\}$; $i \succ_i j$ and let φ be true at i only. Then $M \models_i \mathbf{B}\mathrm{M}\varphi$, since (1) $\exists j M \models_j \varphi$, namely i, and (2) j satisfies $\exists j(Aij \ \&$ $\forall k[(j \succsim_i k \ \& \ Aik) \ \rightarrow \ [M \models_k \neg\varphi \ \& \ \forall l[M \models_l \varphi \ \rightarrow \ l \succ_i j])$, for Aij, and only j itself satisfies the condition on k that $j \succsim_i k \ \& \ Aik$ and it also satisfies $M \models_k \neg\varphi$; lastly, the only φ-world is i and $i \succ_i j$. Nonetheless $M \not\models_i \mathbf{P}\mathrm{E}\varphi$, since $\neg\exists j(Aij \ \& \ M \models_j \varphi)$ for only j satisfies Aij but $M \not\models_j \varphi$. □

Corollary 3.10 $\not\models \boldsymbol{B}M\varphi \ \rightarrow \boldsymbol{B}M^{P^*}\varphi.$[23]

Corollary 3.11 $\not\models \boldsymbol{B}M\varphi \ \rightarrow \boldsymbol{B}M^P\varphi.$[24]

We have thus made conceptual space for *impermissibly* doing more good than the permissible minimum.

Observation 3.2 *The permissibility-entailing supererogation operators are ordered via proper entailment from left to right as* $\boldsymbol{B}M^{P^*}$*,* $\boldsymbol{B}M^P$*,* $\boldsymbol{B}C^P$*. If we include the permissibility-neutral* $\boldsymbol{B}M\varphi$*, the new beyond the permissible minimum operators are ordered in strength left to right as:* $\boldsymbol{B}M^{P^*}\varphi$*,* $\boldsymbol{B}M^P\varphi$*,* $\boldsymbol{B}M\varphi$*, that is,*

$\models \boldsymbol{B}M^{P^*}\varphi \ \rightarrow \boldsymbol{B}M^P\varphi$*;* $\not\models \boldsymbol{B}M^P\varphi \ \rightarrow \boldsymbol{B}M^{P^*}\varphi$*;*[25] $\models \boldsymbol{B}M^P\varphi \ \rightarrow \boldsymbol{B}C^P\varphi$*; and* $\not\models \boldsymbol{B}C^P\varphi \ \rightarrow \boldsymbol{B}M^P\varphi$*;* $\models \boldsymbol{B}M^P\varphi \ \rightarrow \boldsymbol{B}M\varphi$*;* $\not\models \boldsymbol{B}M\varphi \ \rightarrow \boldsymbol{B}M^P\varphi$*.*

Lastly, for $\boldsymbol{B}C^P\varphi$ *and* $\boldsymbol{B}M\varphi$*, neither entails the other. Proofs for these claims is straightforward.*

[22] DWE's $\mathbf{B}\mathrm{C}^{\mathrm{P}}\varphi \overset{def}{=} \mathbf{P}\mathrm{E}\varphi \ \& \ \mathbf{MI}^{\mathrm{P}}\neg\varphi$. But in the model for Proposition 3.5, the truth-conditions for $\mathbf{P}\mathrm{E}(\varphi \vee \psi)$ and $\mathbf{MI}^{\mathrm{P}}\neg(\varphi \vee \psi)$ are satisfied (by i and j respectively), but not so for $\mathbf{B}\mathrm{M}(\varphi \vee \psi)$, as noted for the first corollary.

[23] The first conjunct of the truth condition for $\mathbf{B}\mathrm{M}^{\mathrm{P}^*}\varphi$ is just the truth-condition for $\mathbf{P}\mathrm{E}\varphi$.

[24] By definition, $\models \mathbf{B}\mathrm{M}^{\mathrm{P}}\varphi \ \rightarrow \mathbf{P}\mathrm{E}\varphi$.

[25] We prove $\not\models \mathbf{B}\mathrm{M}^{\mathrm{P}}\varphi \ \rightarrow \mathbf{B}\mathrm{M}^{\mathrm{P}^*}\varphi$ directly and discuss it below (Proposition 3.16)

The next invalidity pair shows that our two maximizing operators are independent; it also illustrates how we can model the tension between classical deontological views and classical optimizing consequentialist views.

Proposition 3.12 *a)* $\not\models \boldsymbol{MA'}\varphi \ \rightarrow \boldsymbol{MA^P}\varphi$ *and b)* $\not\models \boldsymbol{MA^P}\varphi \ \rightarrow \boldsymbol{MA'}\varphi$.

Proof. Let $W = \{i, j\}$, $A^i = \{i\}$, $j \succ i$, $M \models_i \neg\varphi$, and $M \models_j \varphi$. Then
a) $M \models_i \mathbf{M}\mathrm{A}'\varphi$, for j satisfies $\exists j \forall k(k \succsim_i j \ \rightarrow \ M \models_k \varphi)$, but $M \not\models_i \mathbf{M}\mathrm{A}^{\mathrm{P}}\varphi$, that is, $\neg\exists j(Aij \ \&\ \forall k[(k \succsim_i j \ \&\ Aik) \ \rightarrow \ M \models_k \varphi])$, since the only i-acceptable world is i itself and it is ranked as high as itself yet $M \not\models_i \varphi$.
b) Same model except now $M \models_i \varphi$ and $M \models_j \neg\varphi$. Then $M \models_i \mathbf{M}\mathrm{A}^{\mathrm{P}}\varphi$, for i satisfies $\exists j(Aij \ \&\ \forall k[(k \succsim_i j \ \&\ Aik) \ \rightarrow \ M \models_k \varphi])$, since Aii holds and so does $\forall k[(k \succsim_i i \ \&\ Aik) \ \rightarrow \ M \models_k \varphi]$ since only i satisfies the conjunctive antecedent of the quantified conditional and $M \models_i \varphi$; but $M \not\models_i \mathbf{M}\mathrm{A}'\varphi$, since neither i nor j satisfy the condition on j in $\exists j \forall k(k \succsim_i j \ \rightarrow \ M \models_k \varphi)$: i doesn't since $j \succsim_i i$, but $M \not\models_j \varphi$, and j doesn't $j \succsim_i j$ but $M \not\models_j \varphi$. □

Corollary 3.13 $\not\models \boldsymbol{MA'}\varphi \ \rightarrow \boldsymbol{PE}\varphi$.

Note that the first invalidating instance for formula a) in Proposition 3.12 fits with a classic case where you, a surgeon, can save five lives by distributing the organs of one healthy person (sacrificing "the donor") to five others who will die if they don't get them. [26] Here, as above, the classical deontologist can claim that although worlds where five are saved are better than those where none of them are, it is nonetheless impermissible for the agent to access those worlds. Fig 5 pictures the model.

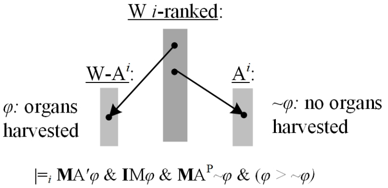

W *i*-ranked:
W-A^i:
A^i:
φ: organs harvested
~φ: no organs harvested
$\models_i \mathbf{M}\mathrm{A}'\varphi \ \&\ \mathbf{I}\mathrm{M}\varphi \ \&\ \mathbf{M}\mathrm{A}^{\mathrm{P}}{\sim}\varphi \ \&\ (\varphi > {\sim}\varphi)$

Fig. 5. Organ Harvesting Deontological Model

Note also that the fact that $\mathbf{M}\mathrm{A}^{\mathrm{P}}$ and $\mathbf{M}\mathrm{A}'$ come apart also raises questions about "ought." Should "ought" be associated with the best per se, or the permissible bests? It seems that the latter must be said if we think that $\mathbf{M}\mathrm{A}^{\mathrm{P}}$ and $\mathbf{M}\mathrm{A}'$ can diverge. We can define such classical deontological dilemmas as follows:

[26] See [21], p.206; but there are multiple cases of this sort such as Trolley cases ([2,22]) and framing the innocent cases [11] to name two others. The cases all circle around the classical Deontologist's central claim: that there are harms "that cannot be justified by the production of a greater good, or the avoidance of a greater harm, for others" ([8], p.464).

Definition 3.14 A set of formulas encodes a classical deontological dilemma in a DWE($\geq$) Model M at a world $i \overset{def}{=}$ there is a formula φ such that: $\models_i \mathbf{M}\mathrm{A}'\varphi$ & $\mathbf{I}\mathrm{M}^{\mathrm{P}}\varphi$.[27]

Let's next model a minor adaptation of a key case of Wessels' for supererogatory holes using this interpretation of three simple formulae:

φ = I donate exactly \$50 (US dollars) and save exactly one life;
ψ = I donate exactly \$5000 and I save exactly 100 lives.
χ = I donate exactly 5000+50 and save exactly 200 lives

For χ, we might imagine a dystopian future with an online fundraiser for which passing a \$5050 threshold triggers another 100 lives saved. Add that \$50 is the permissible minimum and \$50 is of modest utility for me, \$5000 is a very big sacrifice, but \$50 more is just a very slightly bigger sacrifice. Assume as suggested that φ, ψ, χ are mutually exclusive. Let model M be such that:

$$W = \{i, j, k\};\ A^i = \{i, k\};\ M \models_i \varphi,\ M \models_j \psi,\ M \models_k \chi, \text{ and } k \succ_i j \succ_i i.$$

What follows? $\models_i \mathbf{M}\mathrm{I}^{\mathrm{P}}\varphi$ & $\mathbf{B}\mathrm{M}\psi$ & $\mathbf{B}\mathrm{M}^{\mathrm{P}^*}\chi$ & $(\psi > \varphi)$, yet $\models_i \mathbf{I}\mathrm{M}\psi$. Here although we imagine that ψ outranks the minimum, φ, the fact that a very marginally greater donation (adding \$50) than that of ψ provides such a large return in saved lives that it makes ψ impermissible even though ψ is (much) better than the permissible minimum, φ. The "hole" is said to stem from the existence of the still higher ranked accessible alternative χ; χ is supererogatory and it "removes ψ" from the ranking of the *permissible options*—it bumps it into the *impermissibles* bucket, but does not change its relative ranking. This is illustrated in Fig 6 ("k" for thousands).

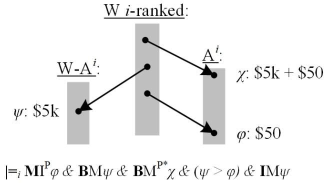

W i-ranked:
W-A^i:
A^i:
χ: \$5k + \$50
ψ: \$5k
φ: \$50
$\models_i \mathbf{M}\mathrm{I}^{\mathrm{P}}\varphi$ & $\mathbf{B}\mathrm{M}\psi$ & $\mathbf{B}\mathrm{M}^{\mathrm{P}^*}\chi$ & $(\psi > \varphi)$ & $\mathbf{I}\mathrm{M}\psi$

Fig. 6. Wessels' Donation Case

Wessels deems ψ to be not supererogatory since not permissible, but she certainly thinks that ψ involves impermissibly doing more good than the permissible minimum. DWE($\geq$) seems to model features of Wessels' case well.[28]

[27] Note that $\models_i \mathbf{M}\mathrm{A}'\varphi \leftrightarrow (\varphi > \neg\varphi)$ by definition, and $\models_i \mathbf{I}\mathrm{M}\varphi \ \rightarrow\ (\neg\varphi >^{\mathrm{P}} \varphi)$.

[28] Since we set aside conditionals here, we do not discuss her argument via conditional operators. See the references in the next note.

We can also use the same model for the *all or nothing problem* scenario, since that scenario is just a special (and especially vivid) case where the marginal sacrifice is the limit of minimal: none. Here we let:

$\varphi =$ I call for help and perform no rescue;
$\psi =$ I rescue exactly one and lose one arm;
$\chi =$ I rescue both and lose one arm.

The intuition of many about this case, as modeled above, is that the minimum involves φ, χ is permissibly beyond the permissible minimum, and ψ yields more good than the permissible minimum but ψ is nonetheless impermissible since it is ruled out by the marginal difference in cost of χ (none) given the large gain. In the framework of DWE($\geq$), once again, $\models_i \mathbf{MI}^{\mathrm{P}}\varphi \ \&\ \mathbf{B}\mathrm{M}\psi \ \&\ \mathbf{B}\mathrm{M}^{\mathrm{P}^*}\chi \ \&\ \psi > \varphi$, yet $\models_i \mathbf{I}\mathrm{M}\psi$, just as in the Wessells' case. [29]

Within our system, we can define supererogatory holes:

Definition 3.15 A set of formulas encodes a supererogatory hole in a DWE($\geq$) Model M at a world $i \overset{def}{=}$ there are formulas φ, ψ, χ meeting these conditions: $\models_i \neg[(\varphi \ \&\ \psi) \vee (\varphi \ \&\ \chi) \vee (\psi \ \&\ \chi)]$ and $\models_i \mathbf{MI}^{\mathrm{P}}\varphi \ \&\ \mathbf{B}\mathrm{M}\psi \ \&\ \mathbf{B}\mathrm{M}^{\mathrm{P}^*}\chi \ \&\ \mathbf{I}\mathrm{M}\psi$. [30]

In Proposition 3.3 we showed that $\models \mathbf{B}\mathrm{M}^{\mathrm{P}^*}\varphi \ \rightarrow \mathbf{B}\mathrm{M}^{\mathrm{P}}\varphi$, but we now show that even though these two new operators are permission-entailing and each also entails the conditions for the **B**M operator, they are nonetheless distinct.

Proposition 3.16 $\not\models \boldsymbol{BM^{P}}\varphi \ \rightarrow \boldsymbol{BM^{p^*}}\varphi$.

Proof. $M : W = \{i, j, k\}$; $A^i = \{i, j\}$; $M \models_i \neg\varphi \ \&\ \neg\psi M \models_j \varphi \ \&\ \neg\psi$, $M \models_k \varphi \ \&\ \psi$, and $k \succ_i j \succ_i i$. Clearly $M \models_i \mathbf{PE}\varphi$, so in particular $\exists j' M \models_{j'} \varphi$. But the second clause for $M \models_i \mathbf{B}\mathrm{M}\varphi$, $\exists j[Aij \ \&\ \forall k((j \succsim_i k \ \&\ Aik) \ \rightarrow [M \models_k \neg\varphi \ \&\ \forall l[(M \models_l \varphi) \ \rightarrow \ l \succ j)])]$, is also satisfied; i satisfies the condition on j, for Aii holds and only i satisfies the antecedent of the universally quantified clause $[i/j, i/k]$ and $M \models_i \neg\varphi$; also $\forall l[(M \models_l \varphi) \ \rightarrow \ l \succ i)]$ holds, since all φ-worlds outrank i. So $M \models_i \mathbf{B}\mathrm{M}\varphi$. But $M \not\models_i \mathbf{B}\mathrm{M}^{\mathrm{P}^*}\varphi$. For its second clause, $\exists j'(Aij' \ \&\ \forall k[(j' \succ_i k \ \&\ Aik) \ \rightarrow \ [M \models_k \neg\varphi \ \&\ \forall l(M \models_l \varphi \ \rightarrow \ (l \succ j' \ \&\ Ail))]])$, fails since the only acceptable $\neg\varphi$-world is i but $\forall l(M \models_l \varphi \ \rightarrow \ (l \succ i \ \&\ Ail)$ fails since k is a φ-world yet Aik fails. □

Here, although $\mathbf{PE}\varphi$ holds and φ guarantees more good than the permissible minimum, there are better worlds where φ and ψ occur but $\mathbf{I}\mathrm{M}\psi$, so that not all cases of φ are unalterably permissible as $\mathbf{B}\mathrm{M}^{\mathrm{P}^*}\varphi$ requires. For ψ here, suppose it involves gains procured by framing the innocent, another prop for a deontological objection to utilitarianism ([11]). Suppose also that my doing

[29] We are abstracting from details by not modeling agent cost and collective or altruistic gain. See [23,24] for Wessel's attempt to do just that and [17] for brief critical exposition.

[30] Note that $\models (\mathbf{B}\mathrm{M}^{\mathrm{P}^*}\chi \ \&\ \mathbf{I}\mathrm{M}\psi) \rightarrow \ (\chi >^{\mathrm{P}} \psi)$ since $\models (\mathbf{PE}\chi \ \&\ \mathbf{I}\mathrm{M}\psi) \rightarrow \ (\chi >^{\mathrm{P}} \psi)$ and $\models \mathbf{B}\mathrm{M}^{\mathrm{P}^*}\chi \ \rightarrow \mathbf{PE}\chi$.

anything to calm the nerves (φ) will do more good than the permissible minimum for me; but if I see to φ by framing an innocent outcast (ψ), I still do more good than doing the bare minimum. This is a case where classical deontologists would say it is sometimes just out of bounds to do the best thing. [31]

In closing this section, let me just note that the above perhaps reflects one advantage of $\mathbf{B}\mathrm{M}^{\mathrm{P}}$ over $\mathbf{B}\mathrm{M}^{\mathrm{P}^*}$. Consider again the cases of supererogatory holes and the all or nothing cases. Let's focus on Wessels' example and now let's add ψ' to the case:

ψ' = I donate at least \$5000 and I save at least 100 lives.

Donating \$5050 (χ) and donating exactly 5000 (ψ) each entail ψ'. What is the status of ψ'? Of course we still have $\models_i \mathbf{MI}^{\mathrm{P}}\varphi$ & $\mathbf{B}\mathrm{M}^{\mathrm{P}^*}\chi$ & $\mathbf{B}\mathrm{M}\psi$, yet $\models_i \mathbf{IM}\psi$, but checking the prior model for the original example, you will see that the following holds:

Both $\mathbf{B}\mathrm{M}\psi'$ & $\mathbf{B}\mathrm{M}^{P}\psi'$ hold, but $\mathbf{B}\mathrm{M}^{\mathrm{P}^*}\psi'$ fails,.

For since $\models\ \chi \rightarrow \psi'$ and $\models_i \mathbf{P}\mathrm{E}\chi$, we get $\models_i \mathbf{P}\mathrm{E}\psi'$, and then $\mathbf{B}\mathrm{M}^{\mathrm{P}}\psi'$ from $\mathbf{B}\mathrm{M}\psi'$, but for the reasons stated regarding the countermodel for Proposition 3.16, we can say of ψ' that not all cases of ψ' are permissible, since ψ isn't. That way of achieving ψ' is ruled out by the strong conditions for $\mathbf{B}\mathrm{M}^{\mathrm{P}^*}$, since $\mathbf{B}\mathrm{M}^{\mathrm{P}^*}\psi'$ requires that *all* worlds where ψ' is realized are *acceptable.* But one might think that donating *at least* \$5000 and saving at least one hundred lives is unproblematically and permissibly beyond the call. For that much of what is done was beyond the call, even if something else stronger, ψ, is impermissibly beyond the call. $\mathbf{B}\mathrm{M}^{\mathrm{P}}$ is a bit more like DWE's $\mathbf{B}$C operator (ψ' is permissible but excluded by doing the minimum), while adding value in that $\mathbf{B}$M's also makes realizing φ *guarantee* exceeding the minimum. We might say of the person who we think wrongfully acts in realizing ψ, "you went permissibly beyond the call in realizing ψ'—that was commendable, but not so for realizing ψ. Similar remarks would apply to the all or nothing case and the status of rescuing *at least* one of the two: realizing *that* was permissibly beyond the call, but not so for rescuing exactly one— that was impermissibly beyond the call. It is good to have each of $\mathbf{B}\mathrm{M}^{\mathrm{P}}$ and $\mathbf{B}\mathrm{M}^{\mathrm{P}^*}$ on the books as well as $\mathbf{B}$M, and like $\mathbf{B}$C, they are all ripe for the picking from the deceptively rich DWE($\geq$) Frames.

4 DWE($\geq$)$^{\mathbf{G}}$ Framework: The Impact of the GOO Constraint

Let's define a special subset of DWE($\geq$) Frames:

[31] Indeed, this model also shows that $\mathbf{MA}'\varphi$ does not guarantee $\mathbf{MA}^{\mathbf{P}}\varphi$ or even $\mathbf{P}\mathrm{E}\varphi$.

Definition 4.1 $F = \langle W, A, \succsim \rangle$, is a DWE$(\geq)^{\mathrm{G}}$ Frame $\overset{def}{=}$ F is a DWE$(\geq)$ Frame such that: $(k \succsim_i j \ \&\ Aij) \ \rightarrow\ Aik$ ["GOO" for "Good as OK is OK"]

Here we merely highlight GOO's significance with a selection of contrasting invalidities and validities. In doing so, "$\models$" stands for validity in DWE$(\geq)$ frames, "$^{\mathrm{G}} \models$" for validity in DWE$(\geq)^{\mathrm{G}}$ frames (similarly for invalidity and "$\not\models$" and "$^{\mathrm{G}} \not\models$"). All proofs are set aside because of space constraints.

Proposition 4.2 $\not\models$ *[*$\boldsymbol{P}E\varphi \ \&\ (\neg\varphi \geq \varphi)] \ \rightarrow\ \boldsymbol{P}E\not\varphi$

Corollary 4.3 $\not\models$ *[*$\boldsymbol{P}E\varphi \ \&\ (\psi \geq \varphi)] \ \rightarrow\ \boldsymbol{P}E\psi$

Proposition 4.4 $^{G} \models$ *[*$\boldsymbol{P}E\varphi \ \&\ (\psi \geq \varphi)$*]* $\rightarrow\ \boldsymbol{P}E\psi$ *(GPP: Good as Permissible is Permissible)*

Proposition 4.5 *a)* $\not\models$ *[*$\boldsymbol{P}E\varphi \ \&\ \boldsymbol{P}E\psi \ \&\ (\varphi > \psi)$*]* $\rightarrow\ (\varphi \geq^{P} \psi)$*;*
b) $\not\models\ \boldsymbol{P}E\psi \ \rightarrow\ [(\varphi \geq^{P} \psi \ \rightarrow\ (\varphi \geq \psi)].$

Corollary 4.6 $\not\models\ [\varphi \geq^{P})] \ \rightarrow\ (\varphi \vee \psi).$

Proposition 4.7 $^{G} \models\ \boldsymbol{P}E\varphi \ \rightarrow\ [(\varphi \geq \psi) \leftrightarrow (\varphi \geq^{P} \psi)].$

Corollary 4.8 $^{G} \models\ (\varphi \geq^{P} \psi) \ \rightarrow\ (\varphi \geq \psi).$[32]

Corollary 4.9 $^{G} \models\ (\varphi \geq^{P} \neg\varphi) \ \rightarrow\ (\varphi \geq \neg\varphi).$

So given GOO, φ ranks as high as ψ among the permissibles only if it ranks as high as ψ, period. What of our maximality operators?

Proposition 4.10 *a)* $\not\models\ \boldsymbol{MA}'\varphi \ \rightarrow\ \boldsymbol{MA}^{P}\varphi$*; b)* $\not\models\ \boldsymbol{MA}^{P}\varphi \ \rightarrow\ \boldsymbol{MA}'\varphi$[33]

Proposition 4.11 $^{G} \models\ \boldsymbol{MA}'\varphi \leftrightarrow \boldsymbol{MA}^{P}\varphi.$

Without GOO, $\mathbf{MA}'\varphi$ and $\mathbf{MA}^{\mathrm{P}}\varphi$ are independent; with GOO, a *permissible best* is just a *best per se*—what I ought to do is the best either way.

Note next an asymmetry in that the minimality operators remain independent, but this is to be expected on reflection:

Proposition 4.12 *a)* $^{G} \not\models\ \boldsymbol{MI}'\varphi \ \rightarrow\ \boldsymbol{MI}^{P}\varphi$*; b)* $^{G} \not\models\ \boldsymbol{MI}^{P}\varphi \ \rightarrow\ \boldsymbol{MI}'\varphi$

Corollary 4.13 *a)* $\not\models\ \boldsymbol{MI}'\varphi \ \rightarrow\ \boldsymbol{MI}^{P}\varphi$*; b)* $\not\models\ \boldsymbol{MI}^{P}\varphi \ \rightarrow\ \boldsymbol{MI}'\varphi$

What of the operators for exceeding the permissible minimum? Consider these contrasts:

Proposition 4.14 *a)* $\not\models\ \boldsymbol{B}M\varphi \ \rightarrow\ \boldsymbol{B}M^{P}\varphi$*; b)* $\not\models\ \boldsymbol{B}M^{P}\varphi \ \rightarrow\ \boldsymbol{B}M^{P^{*}}\varphi$

Proposition 4.15 $^{G} \models\ \boldsymbol{B}M\varphi \ \rightarrow\ \boldsymbol{B}M^{P^{*}}\varphi$

Corollary 4.16 $^{G} \models\ \boldsymbol{B}M\varphi \ \rightarrow\ \boldsymbol{P}E\varphi$[34]

Corollary 4.17 $^{G} \models\ \boldsymbol{B}M\varphi \leftrightarrow \boldsymbol{B}M^{P}\varphi$[35]

Corollary 4.18 $^{G} \models\ \boldsymbol{B}M\varphi \leftrightarrow \boldsymbol{B}M^{P^{*}}\varphi$[36]

[32] The converse is invalid with or without GOO, since φ and ψ might both be impermissible.

[33] We proved this in Section 3.2 as Proposition 3.12 there.

[34] Recall the proof in Section 3.2 that $\models\ \mathbf{B}\mathrm{M}^{\mathrm{P}^{*}}\varphi \ \rightarrow\ \mathbf{P}\mathrm{E}\varphi$

[35] Recall $\mathbf{B}\mathrm{M}^{\mathrm{P}}\varphi \overset{def}{=} \mathbf{B}\mathrm{M}\varphi \ \&\ \mathbf{P}\mathrm{E}\varphi$

[36] Recall the proof in Section 3.2 that $\models \mathbf{B}\mathrm{M}^{\mathrm{P}^{*}}\varphi \ \rightarrow\ \mathbf{B}\mathrm{M}\varphi$

Corollary 4.19 ${}^{G}\models\ \boldsymbol{B}M^{P}\varphi \leftrightarrow \boldsymbol{B}M^{P^*}\varphi$

So, among other reductions, our three new beyond-the-minimum operators are equivalent given GOO, and the *conceptually* permission-neutral BM operator is now permissibility entailing. To go beyond the call at all is to do so *unalterably* in a *permissible manner* that *guarantees* more *good* than the minimum. Unsurprisingly, **B**C from DWE remains distinct from these three.

Proposition 4.20 ${}^{G}\not\models \boldsymbol{BC}\varphi\ \rightarrow\ \boldsymbol{BM}\varphi$

Corollary 4.21 ${}^{G}\not\models \boldsymbol{BC}\varphi\ \rightarrow\ \boldsymbol{BM}^{P}\varphi$

Corollary 4.22 ${}^{G}\not\models \boldsymbol{BC}\varphi\ \rightarrow\ \boldsymbol{BM}^{P^*}\varphi$

Given GOO, even the *conceptually* permission-neutral **B**M operator is now stronger than **B**C:

Proposition 4.23 ${}^{G}\models \boldsymbol{BM}\varphi\ \rightarrow\ \boldsymbol{BC}\varphi$

What about the indifference notions?

Proposition 4.24 $\not\models \boldsymbol{VI}\varphi\ \rightarrow\ \boldsymbol{IN}^{P}\varphi$

Corollary 4.25 $\not\models \boldsymbol{VI}\varphi\ \rightarrow\ \boldsymbol{CI}^{P}\varphi$

Proposition 4.26 ${}^{G}\models \boldsymbol{VI}\varphi\ \rightarrow\ \boldsymbol{IN}^{P}\varphi.$

Corollary 4.27 ${}^{G}\models \boldsymbol{CI}\varphi \leftrightarrow \boldsymbol{VI}\varphi$ *(Recall* $\boldsymbol{CI}\varphi \overset{def}{=} \boldsymbol{VI}\varphi\ \&\ \boldsymbol{IN}\varphi$*)*

Finally, we have:

Observation 4.1 *Given GOO, there are no deontological dilemmas per Definition 3.14 (for* $\boldsymbol{MA'}\varphi$ *now entails* $\boldsymbol{MA}^{P}\varphi$*, and so* $\boldsymbol{PE}\varphi$*) nor supererogatory holes per Definition 3.15 (for* $\boldsymbol{BM}\varphi$ *now entails* $\boldsymbol{BM}^{P^*}\varphi$*, and so* $\boldsymbol{PE}\varphi$*).*

All these results reflect the fact that given GOO, the ordered worlds per i (along with a subset selection, A^{i}) are just the ordered i-acceptables stacked on top of the ordered i-unaceptables, as in Figure 7.

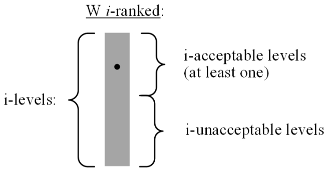


Fig. 7. Impact of GOO Constraint

We have established that GOO is a *very significant* principle for deontic logic, as is its analogue for ethics, GPP.[37] It is plainly very reductive and results in substantial theoretical simplifications.

[37] Likewise for GPP: what is as good as permissible is permissible.

5 Brief Concluding Remarks

Space limitation constrained us from exploring other semantic features, and from exploring what logics for the DWE($\geq$) and DWE($\geq$)$^{\mathrm{G}}$ frameworks might look like, as well as exploring other operators we might have defined (e.g. conditionalized versions of some of the operators). [38] There is also a way to *generate* the framework above for DWE($\geq$)$^{\mathrm{G}}$ from a more fundamental framework in a substantively plausible way (a hybrid deontological cum consequentialist framework (see [1] for general background on hybrids). Recasting things in a less classical framework would be worthwhile as well. For example, the semantic ordering relation needn't be connected. We leave these and other changes for the future.

I hope we've made clear that classical deontological dilemmas are cousin to the problem of supererogatory holes: for each hinges on a longstanding but growing divide in normative ethics over whether something as good as something permissible must be permissible. [39] We also saw that the all or nothing case is a special case of Wessels' supererogatory holes. The expanded DWE framework allows us to inherit some of the advantages of the original when it comes to reconsidering supererogatory holes. First, focusing on the latter case, even if we deny the permissibility of saving just one (ψ), we can still get the result that saving at least one (ψ') was permissibly beyond the call (i.e. $\mathbf{B}\mathrm{M}^{\mathrm{P}}\psi'$) even if de facto done impermissibly by saving exactly one ($\mathbf{B}\mathrm{M}\psi$ & $\mathbf{I}\mathrm{M}\psi$). This is a byproduct of solving DWE's Problem of Disjunctive Supererogation. Secondly, suppose we embrace GOO and so deny the possibility of impermissibly going beyond the call? The resources we have naturally allow us to still say unequivocally that Jane *ought to not* rescue just one ($\mathbf{M}\mathrm{A}^{\mathrm{P}}\neg\psi$) and that certainly seems to take some of the alleged sting out of just saying it is ok for Jane to save just one. We can say she shouldn't do that, even if it is not the case that she must not ($\neg\mathbf{O}\mathrm{B}\neg\psi$). It is important to clearly distinguish *must* from the weaker *ought.* [40] If we do say morality forbids rescuing just one, then mustn't we say of Jane who rescues just one, losing an arm in the process "*The least Jane could have done* was just call for help"? That sounds unpalatable. One could, I guess, get rid of the supererogatory hole by saying that in not saving the other person when it could be done with no more cost than saving just one that is so unjust that it is outright worse than just going for help (saving none), but I find that hard to swallow. It seems more plausible to say that (morally speaking) Jane should not go for help, and she should not rescue just one (even though doing so is permissibly beyond the call of duty); what she (and all of us) *ought* to do is even better than that: rescue both.

[38] [18] has made a start in this direction.

[39] Hansson's impressive [5] is a noteworthy prior case in deontic logic where the GPP principle is put to important use. GOO was also put to more restricted use in [18].

[40] As argued for in detail in [13], where the first model theoretic account is offered of the difference.

References

[1] Alexander, L., *Deontological Ethics*, in: E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, 2021, Winter 2021 edition . URL
`https://plato.stanford.edu/archives/win2021/entries/ethics-deontological/`

[2] Foot, P., *The problem of abortion and the doctrine of the double effect*, Oxford Review **5** (1967), pp. 5–15.

[3] Goble, L., *A logic of good, should, and would: Part I*, Journal of Philosophical logic (1990), pp. 169–199.

[4] Goble, L., *A logic of good, should, and would: Part II*, Journal of Philosophical Logic (1990), pp. 253–276.

[5] Hansson, S. O., "The structure of values and norms," Cambridge University Press, 2001, 314 pp.

[6] Hansson, S. O., *Representing supererogation*, Journal of Logic and Computation **25** (2013), pp. 443–451.

[7] Horton, J., *The all or nothing problem*, The Journal of Philosophy **114** (2017), pp. 94–104.

[8] Huemer, M., *A paradox for weak deontology*, Utilitas **21** (2009), pp. 464–477.

[9] Krogh, C. and H. Herrestad, *Getting personal some notes on the relationship between personal and impersonal obligation*, in: M. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems*, Springer Verlag: New York, 1996, pp. 134–153.

[10] Mares, E. D. and P. McNamara, *Supererogation in deontic logic: Metatheory for DWE and some close neighbours*, Studia Logica **59** (1997), pp. 397–415.

[11] McCloskey, H. J., *An examination of restricted utilitarianism*, The Philosophical Review **66** (1957), pp. 466–485.

[12] McNamara, P., *Making room for going beyond the call*, Mind **105** (1996), pp. 415–450.

[13] McNamara, P., *Must I do what I ought?(or will the least I can do do?)*, in: M. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems*, Springer Verlag: New York, 1996, pp. 154–173.

[14] McNamara, P., *Agential obligation as non-agential personal obligation plus agency*, Journal of Applied Logic **2** (2004), pp. 117–152.

[15] McNamara, P., *Praise, blame, obligation, and dwe: Toward a framework for classical supererogation and kin*, Journal of Applied Logic **9** (2011), pp. 153–170.

[16] McNamara, P., *Supererogation, inside and out: Toward an adequate scheme for common sense morality*, in: M. Timmons, editor, *Oxford Studies in Normative Ethics*, I (2011).

[17] McNamara, P., *Logics for supererogation and allied concepts*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems, Volume 2* (2013).

[18] McNamara, P., *A natural conditionalization of the DWE framework*, in: P. McNamara, Brown, Mark, Jones and Andrew, editors, *Agency, Norms, Inquiry, and Artifacts: Essays in Honor of Risto Hilpinen*, Springer Nature, Synthese Library 454 Switzerland, 2022 pp. 113–136.

[19] Muñoz, D., *Three paradoxes of supererogation*, Noûs **55** (2021), pp. 699–716.

[20] Pummer, T., *All or nothing, but if not all, next best or nothing*, The Journal of Philosophy **116** (2019), pp. 278–291.

[21] Thomson, J. J., *Killing, letting die, and the trolley problem*, The Monist **59** (1976), pp. 204–217.

[22] Thomson, J. J., *The trolley problem*, Yale Law Review **94** (1985), pp. 1395–1415.

[23] Wessels, U., "Die gute Samariterin Zur Struktur der Supererogation," Berlin, de Gruyter, 2002.

[24] Wessels, U., *Beyond the call of duty: The structure of a moral region*, in: C. Cowley, editor, *Supererogation*, Cambridge University Press, Cambridge, 2015 pp. 87–104.

On Floating Conclusions

Daniela Schuster [1]

Universität Konstanz

Jan Broersen [2]

Utrecht University

Henry Prakken [3]

Utrecht University, University of Groningen

Abstract

When there are two lines of argument that contradict each other but still end up with the same conclusion, this conclusion is called a floating conclusion. It is an open topic in skeptical defeasible reasoning if floating conclusions ought to be accepted. Interestingly, the answer seems to be changing for different examples. In this paper, we propose a solution for explaining the different treatments of the floating conclusion in the various examples from the literature. We collect the examples from the literature, extend them with additional examples and test various hypotheses for explaining the difference by means of the examples. We will argue for a framework that accepts a floating conclusion by default but allows for reasons to deviate from the default in order to reject it. These reasons nicely explain the different underlying patterns of our intuitions.

Keywords: Floating Conclusions, Defeasible Reasoning, Skeptical Reasoning

Acknowledgement

Support by the focus area on Human-centered Artificial Intelligence at Utrecht University, Studienstiftung des deutschen Volkes, Baden-Württemberg Stiftung is gratefully acknowledged. We would like to thank the members of the Logic Reading Group at Utrecht University and three anonymous reviewers for their helpful feedback.

[1] daniela.schuster@uni-konstanz.de

[2] j.m.broersen@uu.nl

[3] h.prakken@uu.nl

1 Introduction

Floating conclusions are a phenomenon that appears in the context of defeasible or non-monotonic reasoning. It was investigated already early in [6] and [7]. When there are (at least) two lines of reasoning that contradict each other, but still end up with the same conclusion, this conclusion is called a floating conclusion. One famous example is the Nixon case. In this example, we have one line of reasoning starting from the fact that Nixon is a republican from which it can be concluded (defeasibly) that Nixon is a hawk from which again it can be concluded that Nixon is politically extreme. A different line of reasoning starts from the fact that Nixon is a quaker from which it can be concluded that Nixon is a dove from which again it can be concluded that Nixon is politically extreme. These two lines of reasoning contradict each other, because Nixon cannot be both a hawk and a dove. We have to reject one line of reasoning. Still, both lines of reasoning, albeit conflicting, end up with the same conclusion: the floating conclusion that Nixon is politically extreme. Should we accept this floating conclusion then after all? This is the question that immediately arises and that is going to be the topic of this paper. The name 'floating conclusion' that stems from [7] nicely captures that the conclusion 'floats' above the conflicting arguments. The question of whether we should accept floating conclusions is tied closely to the question of whether we should accept at least one line of reasoning among a set of conflicting lines of reasoning. This builds on the intuition that all the reasoning lines involved have, albeit being fallible, a certain credibility or plausibility. When a conflict between them arises, it becomes clear that at least one line of reasoning fails at some point. Given that it is not clear which line of reasoning fails, we cannot simply accept one and reject the other. However, can we still assume that *there is* (at least) one line of reasoning that is sound? If this is the case, then we should accept a floating conclusion. If this is not the case, then we should not.

Interestingly, there is not one clear answer to this question. In different examples of floating conclusions, we seem to have contradicting intuitions about whether the floating conclusion should be accepted or not. In other words: in conflicting situations, we sometimes think that at least one line of reasoning is sound, while at other times we think that the conflict between the lines of reasoning destroys *both* conflicting reasoning lines. Floating conclusions are one of the most exciting phenomena in the area of defeasible reasoning, but they also pose an unresolved problem in terms of how to deal with them. Therefore, floating conclusions expose possible imitations in defeasible reasoning and also in automated decision-making. A systematic treatment of floating conclusions is missing in the literature so far. Especially in deontic contexts, this can resolve in an alarming inability to derive norms of action in certain situations.

In this paper, we provide a systematic treatment of the phenomenon of floating conclusions. Thereby, we aim to explain the different intuitions concerning the acceptability of floating conclusions in the different examples. It is important to note here that our approach is based on intuitions. We try to provide a theory that manages to explain pre-theoretic intuitions about different examples and situations. This method is not undisputed. As it has been noted in [12] and again in [9] the use of intuitions in logic has at least two difficulties. One difficulty is the question of whose intuitions should count (as people might differ in their intuitions). The second difficulty questions the assumption that intuitions should always be taken at face value. In fact, Veltman [12, p. 10] argues that when looking for intuitions, we are usually interested in the pre-theoretic judgments of 'common people' who are no experts in the field and have not been exposed to theories about the topic yet. However, then we cannot tell whether these judgments represent knowledge or barely some kind of fallible belief. Hence, those judgments are fallible and do not provide a "rock bottom empirical basis for testing logical theories" [12, p. 13]. Moreover, (good) theories and arguments can surely guide and change intuitions and pre-theoretic judgments [12, p. 15]. Hence, it is important to not blindly rely on any intuition. Nevertheless, theories should also not contradict all broadly accepted 'common-sense' judgments. Although different people may differ in their intuitions about the acceptability of one floating conclusion or the other, there is clear empirical evidence that some floating conclusions are commonly regarded as acceptable, while others clearly are not (especially when it is considered that people should *act* on these conclusions). A theory that accepts all floating conclusions is just as unsatisfactory as a theory that rejects all floating conclusions. In this paper, we do not blindly rely on any intuition. Rather, as Prakken [9] already suggests, we are searching for some underlying pattern in (commonly shared) intuitions and thereby try to explain similarities and differences.

First, we will present several different examples of non-monotonic arguments that involve a floating conclusion. Many examples are discussed in the literature already, others have been constructed for this paper specifically. We will see that the different examples trigger different intuitions about whether we should accept the respective floating conclusion. Next, we will present different hypotheses that try to explain these conflicting intuitions and we will test the validity of these hypotheses with the help of our examples. After having tested the different hypotheses, we will argue that there is not one single explanation that manages to explain all the different intuitions. Instead, our presented solution will take some ingredients from different explanations. We argue that, per default, floating conclusions are to be accepted. However, there are reasons to deviate from the default and to reject a floating conclusion. We will present two different reasons for deviation that together nicely explain and cover all the presented examples. One reason applies if there is a possible 'compromise' between the conflicting elements of the arguments; the second reason applies if the conflict is harmful not only to the conflicting part of the argument but also to other non-conflicting parts because the conflict undermines the credibility of the sources of information altogether. Both of these reasons are based on the fact that in situations of floating conclusions, the conflicting propositions are *contrary.*[4] This means that the propositions cannot be true together, but yet can be false together. The two explanations which give us reason to deviate from the default both spell out a way in which the conflicting propositions are both false, offering a third alternative beyond the two (in the arguments displayed) alternatives that one proposition is true and the other one false or vice versa.

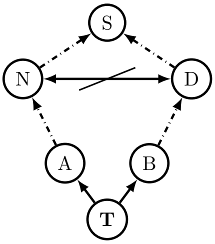
2 Examples of floating conclusions

We will present examples of arguments that involve a floating conclusion. Some of the examples can be found in the literature, others are invented for this paper in order to obtain a precise impression of the phenomenon that is as comprehensive as possible. In a second step, we will divide them by means of the different intuitions about the acceptance of the respective floating conclusion.

2.1 Presenting the examples

In the following, we will use capital letters to abbreviate the sentences or propositions. The arguments are visualized via arrows connecting the sentences. The non-dashed arrows represent strict, monotonic reasoning, while the dashed arrows represent defeasible inferences.[5] The double-sided crossed-out arrow visualizes a conflict between two sentences, while the **T** stands for 'truth.' Sentences that follow from **T** are taken to be known.

Ice-Skating [9]



A Brigt Rykkje has a Norwegian name.
B Brigt Rykkje was born in the Netherlands.
N Brigt Rykkje is Norwegian.
D Brigt Rykkje is Dutch.
S Brigt Rykkje likes ice-skating.

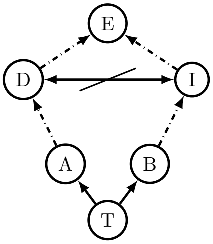
The argument that is visualized by the picture then reads as follows: It is both true (hence strictly follows from (**T**)) that Brigt Rykkje was born in the Netherlands (B) and that he has a Norwegian name (A). The argument on the right side tells us that Brigt Rykkje is Dutch (D) since he was born in the Netherlands (B). The argument on the left tells us that Brigt Rykkje is Norwegian (N) since he has a Norwegian name (A). These two statements, (N) and (D), however, contradict each other and cannot be both true at the same

[4] Thanks to Michael De for pointing us towards this.

[5] One could also call them material inferences in the terminology of [11] and [1].

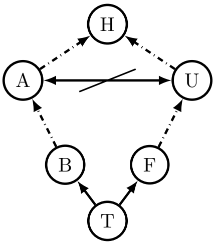
time.[6] On the right side, the argument continues: Brigt Rykkje likes ice-skating, since he is Dutch (D). On the left side, the argument continues: Brigt Rykkje likes ice-skating (S), since he is Norwegian (N). Hence, both argument lines end up with the floating conclusion that Brigt Rykkje likes ice-skating (S). All reasoning steps here are beliefs.

Economy [4]



A Economist A says we will have deflation.
B Economist B says we will have inflation.
D We will have deflation.
I We will have inflation
E We will have economic downturn.

Student Housing [2]

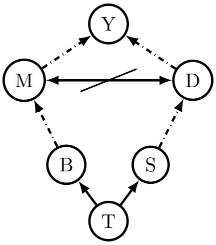


B Susan's boyfriend studies in Amsterdam.
F Susan's best friend studies in Utrecht.
A Susan wants to study in Amsterdam.
U Susan wants to study in Utrecht.
H It will be difficult to get housing.

Here not only beliefs but also desires are involved. For example: Susan *wants* to study in Amsterdam (A), because her boyfriend studies in Amsterdam (B). But again, she *believes* that housing will be very expensive if she studies in Amsterdam.

[6] In almost all the examples, we are making some empirical assumptions, like here: It is not possible to have two citizenships. This reflects the fact that we are reasoning in a non-monotonic setting with incomplete knowledge. One could say that the inferences we draw are material inferences rather than formal inferences in the terms of Sellars [11] and Brandom [1].

Yacht [4]



B My brother tells me that dad will give his half a million dollars to him, but mom will give it to me.

S My sister tells me that mom will give her half a million dollars to her, but dad will give it to me.

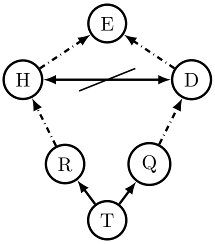
M I will get half a million dollars from my mom.

D I will get half a million dollars from my dad.

Y I put a high deposit on a Yacht that costs half a million dollars.

This example from Horty [4] is about a situation where I have a brother and sister. Our parents are separated and will both die soon. The parents each have a fortune of half a million dollars. Before both parents went into comas, my brother talked to my father and my sister talked to my mother. My sister tells me that according to my mother, my mother will give her half a million dollars to her (my sister), but my father will give his half a million dollars to me. My brother tells me that according to my father, my father will give his half a million dollars to him (my brother), but my mother will give her half a million dollars to me. In this story, I really want to buy a (very particular) yacht for half a million dollars and I intend to make a very large down payment on the yacht should I receive half a million after my parents die.

Nixon [4]



R Nixon is a republican.

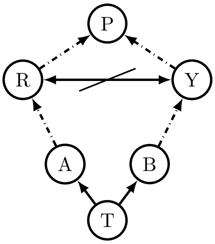
Q Nixon is a quaker.

H Nixon is a hawk.

D Nixon is a dove.

E Nixon is politically extreme.

Primary Color



A Anna says that the cup is red.

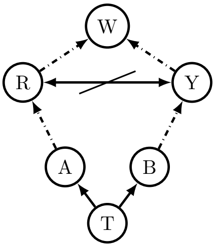
B Ben says that the cup is yellow.

R The cup is red.

Y The cup is yellow.

P The cup is colored in a primary color.

Wavelength Color



A Anna says that the cup is red.

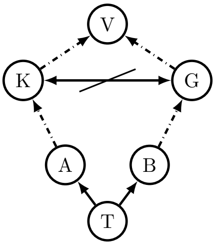
B Ben says that the cup is yellow.

R The cup is red.

Y The cup is yellow.

W The color of the cup has a higher wavelength than the wavelength of blue.

Murderer [9]



A Witness A says that Peter killed the victim with a knife.

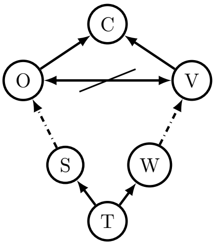
B Witness B says that Peter killed the victim with a gun.

K Peter killed the victim with a knife.

G Peter killed the victim with a gun.

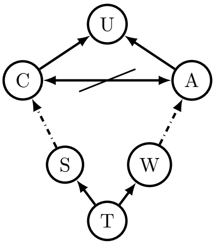
V Peter killed the victim.

Mary in Canada [10] and [4]



S Mary's spouse lives in Toronto.
W Mary's work is in Vancouver.
O Mary lives in Toronto.
V Mary lives in Vancouver.
C Mary lives in Toronto or in Vancouver.

Carol in the US [4]



S Carol's spouse lives in College Park.
W Carol's work is in Alexandria.
C Carol lives in College Park.
A Carol lives in Alexandria.
U Carol lives in College Park or in Alexandria.

2.2 Intuitions about the acceptance of the floating conclusions in the examples

Floating conclusion accepted	Floating conclusion rejected
Ice-Skating	**Economy**
Student Housing	**Yacht**
Wavelength Color	**Primary Color**
Mary in Canada	**Carol in the US**
	Nixon
	Murderer

The table shows the examples in which the floating conclusion should be intuitively accepted and the examples in which it should not. As mentioned in the introduction, the use of intuitions raises some questions. Of course, some people's intuitions may diverge from the table. Although we have not conducted a scientific study, we have asked enough people about their intuitions regarding these examples to assume that the table is representative. Of course, we are also aware that most of the people we interviewed come from an academic background and that the intuitions of other groups of people might be different.

The Nixon case is probably the most controversial and therefore the most interesting case. In former literature [3] people argued that the floating conclusion in the Nixon case should be accepted. However, we think that it should be rejected. Especially when we explain our reasons *why* the Nixon floating conclusion should be rejected, people seem to sometimes change their intuitions and admit that in fact, one cannot conclude that Nixon is politically extreme. Without presuming additional knowledge about Nixon as a person, it is natural to think that his quaker and his republican side 'balance each other out' such that he ends up with no politically extreme stance. This reflects nicely what we said in the introduction. Intuitions are not infallible and they are not necessarily stable. Sometimes good explanations can change intuitions.

3 Hypotheses

In this section, we present different hypotheses that aim to explain why some but not all floating conclusions seem acceptable. We do not claim that the list of hypotheses is exhaustive, nor that all of the hypotheses are prima facie equally convincing. The list contains the hypotheses we found in the literature so far and new hypotheses that we took to be reasonable and worth mentioning. In the subsequent subsection, we will then evaluate the hypotheses by virtue of our examples.

3.1 Presenting the hypotheses

(i) **Vagueness**: One possible explanation is bound to the concept of vagueness. Some conflicts can be seen as borderline cases for vague concepts that are involved in the corresponding defaults. If a vague concept is involved, and the conflicting propositions incorporate a clear, non-borderline case of the concept, it has to be tested whether the floating conclusion also follows from the borderline case. If the floating conclusion does not follow from the borderline case, it should be rejected.

(ii) **The direction of fit** [2]: The difference could stem from different direction of fits. Beliefs are propositions that aim to describe the world, hence the direction of fit can be described as proposition-to-world. Desires and intentions, on the other hand, are propositions that describe how the world ought to be, so the direction of fit is world-to-proposition. This is why conflicting beliefs 'cancel each other out,' resulting in the rejection of the floating conclusion. Conflicting desires or intentions, on the other hand, do not cancel each other out. Thus, at least one of the desires will remain intact and the floating conclusion is to be accepted.

(iii) **Hidden Defaults** [9]: This explanation states that the reason that some floating conclusions might seem unacceptable, results from implicit "hidden defaults" that are not mentioned explicitly, but have to be thought along the respective examples. These hidden defaults defeat (through undercutting) the presented defaults that lead to the alleged floating conclusions, such that these in fact are no floating conclusions but conclusions of defeated defaults.

(iv) **Possible Compromise**: This explanation suggests that one has to look at the compatibility of the conflicting propositions. If there is a possible 'compromise,' or intermediate position, between the conflicting propositions, it is likely that this compromising position is in fact the case. In such a situation, one has to check if the floating conclusion also follows from the compromising case. If it follows *only* from the presented 'extreme' cases but not from the compromising one, the floating conclusion must be rejected. If there is no compromise between the conflicting propositions, it is justified to think that at least one of the conflicted propositions is true, and hence that the floating conclusion is acceptable.

(v) **Harmfulness of the conflict**: This explanation takes a closer look at the conflict, as well as at the sources of information. Sometimes it seems that the conflict is only harmful to the conflict itself. In other cases, though, the conflict seems to destroy the credibility of the sources of information more generally. If this is the case, there is no longer a reason to assume that at least one line of reasoning is sound which results in rejecting the floating conclusion.

3.2 Testing the hypotheses

In this section, we will test the presented hypotheses by means of our examples. We will see that, while most hypotheses manage to explain certain examples well, no hypothesis manages to explain the intuitions behind every example presented.

(i) **Vagueness:** The vagueness hypothesis is motivated by examples like **Wavelength Color** and **Primary Color**. These two examples involve a vague concept (a color). Clearly, the conflicting propositions (that the cup is red and that the cup is yellow) can be dissolved by a third proposition (that the cup is orange) representing the borderline case. In the **Primary Color** example, the floating conclusion does not follow from the borderline case (thus the conclusion is rejected), for **Wavelength Color** the floating conclusion does follow from the borderline case as well (thus it is accepted). The involvement of vagueness alone does not do the job of explaining the differences though. Moreover, there are plenty of examples that do not involve vagueness and for which we still have varying intuitions. These cannot be explained by this hypothesis. However, it becomes clear quite easily that vagueness alone cannot explain all examples. For example, there is **Yacht** which is a rejection example and **Ice-Skating** which is an acceptance example, but neither of the examples involves a vague concept.

(ii) **The direction of fit:** The idea that a different direction of fit can lead to different intuitions about the acceptability of floating conclusions was originally motivated in [2] by the different intuitions in the examples **Economy** and **Student Housing**. In the latter example, the conflict arises due to conflicting desires. Susan wants to study both in Utrecht and in Amsterdam. Although it is clear, that one desire will 'defeat' the other eventually, the desires do not cancel each other out as in the Economy case where we have a conflict between beliefs. However, this explanation fails in other examples. **Ice-Skating** is an example that is free of desires and intentions and purely based on beliefs. Still, we want to accept the floating conclusion in **Ice-Skating**. [7]

(iii) **Hidden Defaults:** [9] argues that the examples **Yacht** and **Murderer** do in fact not provide a reason to reject floating conclusions. The propositions that conflict each other and from which the floating conclusions follow are in both cases defeated since the defaults leading to these conclusions are undercut by other defaults, that are not mentioned explicitly in the theory. In the case of **Murderer**, what makes the alleged floating conclusion unacceptable is the hidden default that, if two witnesses say contradicting things, their credibility is dismissed. This default then undercuts both the default that concludes that Peter killed the victim with a gun and the default that concludes that Peter killed the victim with a knife, yielding no floating conclusion. Likewise in the **Yacht** example, a hidden default will undercut both arguments that rely on the testimonies of my sister and my brother. [8] This strategy succeeds in other examples as well. In the Nixon case, one could find an additional, hidden default stating that if someone is both a quaker and a republican, one cannot tell anything about his or her opinion with respect to military operations. This hidden default would then undercut both defaults that infer either that Nixon is a dove or that Nixon is a hawk. The floating conclusion that he is politically extreme would then not follow either. The rather clear case of **Ice-Skating** also speaks in favor of this hypothesis. There is no apparent hidden default that should be visualized in the example, leading to the intuitively correct conclusion that the floating conclusion is acceptable. The strategy, however, becomes more questionable when examples like **Mary in Canada** and **Carol in the US** (or the Color examples)

[7] One can easily see how the explanation fails in the other direction as well. If one adapted **Carol in the US** to an example about Carol's desires to live in one and the other city, the hypothesis would state that the floating conclusion is to be accepted, although we want to reject it.

[8] Note that Prakken [9] described the example slightly differently. In his description, both my sister and my brother tell me that they spoke to both parents and that my mom (respectively my dad) told my sister (respectively my brother) that she will give me her (his) money. Prakken argues that this example relies on the additional default that people tend to speak the truth about their intentions, which is undercut as soon as people (in this case both mom and dad) tell conflicting things about their intentions.

are considered, where the same defaults in one case lead to seemingly acceptable floating conclusions and in another case to unacceptable ones. Why should there be hidden defaults in one case but not in the other? Prakken himself also admits that this strategy might not be valid for all possible examples, such as conflicts due to different interpretations of legal norms. Moreover, we think that, although this thesis might be applicable for a lot of examples, it does not really provide an explanation about *why* in certain situations a floating conclusion is to be accepted and in others not. By referring only to possible missing defaults, we might get a way out of the unequal treatment of the different floating conclusions, but it still shifts the burden of explanation only to the question about why we feel like there are some defaults missing (or hidden) in some cases, while in other cases this is not so.

(iv) **Possible Compromise:** The idea behind this thesis can best be visualized by the different intuitions of the **Mary in Canada** and **Carol in the US** case. Although the defaults leading to the conflict and to the floating conclusions are of the exact same form, the floating conclusion seems justified in one case and not in the other (as [4] notices.) What explains the difference in this particular case? It seems like the conclusion that Mary lives either in Vancouver or in Toronto is acceptable because there is not really an alternative option in the 'middle.' Since the cities are extremely far away from each other, it is not likely that Mary could live somewhere in the middle and commute between the places on a daily basis. This is different in the case of Carol in the US. Since both cities, College Park and Alexandria, are in fact not very far away from each other and there is a good 'compromise,' Washington D.C., that is in the middle, it is likely that Carol neither lives in College Park nor in Alexandria, but went for the compromise, the city in between. This idea can be transferred to other examples, too. In the **Economy** case, there is a 'compromise'[9] between (strong) deflation and inflation, namely that there will be none of both. Likewise in the **Nixon** case, the compromise between Nixon being a Hawk and Nixon being a Dove lies clearly in the middle in describing Nixon as not having a clear or extreme opinion on military use. In both cases, we do not want to accept the floating conclusion, because the compromise is just too likely and from the compromise the floating conclusion does

[9] Note that the use of the word 'compromise' may be somewhat unusual in this context. Not in all the cases described is there really a compromise in the sense of people agreeing on something. What we mean here by compromise is rather an unignorable possibility or relevant alternative. We use the word 'compromise' anyway because it suggests so nicely that this alternative or possibility lies somewhere *in the middle* on a spectrum at the end of which the two conflicting options lie (and is not simply some additional alternative that lies outside the spectrum considered so far).

not follow.[10] This is different in the cases of **Ice-Skating** and **Student Housing**. There is no attractive student town between Amsterdam and Utrecht and even less is it possible, that Brigt Rykkje can have a citizenship 'in between' Norwegian and Dutch. Therefore, we should stick with the conclusion that, even if he cannot have both, he has at least one of the citizenships, such that the floating conclusion can be drawn. The general idea is that if there is no compromise between the conflicting propositions, then it is likely that at least one line of reasoning is correct and the floating conclusion will follow. If there is a plausible compromise, then it has to be tested if the floating conclusion follows from this compromise, too. This can be nicely visualized via the two color examples. In the identically constructed examples **Wavelength Color** and **Primary Color**, the compromise (that the cup is orange) entails one floating conclusion (that the cup is colored with a higher wavelength than the one of blue) but not the other floating conclusion (that the cup is colored in a primary color). However, the examples of **Murderer** and **Yacht** cannot be perfectly explained by this hypothesis. The reason why we want to reject the floating conclusion is not that there seems to be a compromise or intermediate position between the two conflicting propositions. Rather, it seems like the sole fact that there *is* a conflict undermines the credibility of both argument lines.

(v) **Harmfulness of the conflict:** The thesis about the harmfulness of the conflict is based exactly on this observation concerning **Yacht** and **Murderer**. The basic idea is that there are different kinds of conflicts. Some kinds of conflicts are harmful to the floating conclusion, others are not. The cases of **Yacht** and **Murder**, for example, involve a conflict in which two witnesses assess different things that, although in conflict with each other in some respect, are consistent with each other in another respect. In the Murderer case the witnesses' testimonies conflict in respect to the murder weapon they describe Peter to have used, but they agree upon the fact that it was Peter who killed the victim. In the case of the Yacht, the siblings' testimonies are in conflict with each other in respect to what Mom and Dad will do with their half a million dollars, but they agree that I will end up having half a million dollars from one of them. Still, we wouldn't want to conclude that Peter killed the victim or that I will inherit half a million dollars. Why is this? The conflict involved seems to be harmful not only to the conflicting part itself but harmful to the whole situation as such. The existence of the conflict puts us in doubt about the credibility of the witnesses and makes us suspect that something more general 'has gone wrong.' We might suspect that the two witnesses or the siblings have arranged their statements, or that the conditions for seeing

[10] Horty [4, p. 69] already suggests something similar in his considerations of **Economy** and **Nixon**: "Perhaps the extreme predictions are best seen as undermining each other and the truth lies somewhere in between."

Peter kill the victim weren't that great or that our parents have no intention to reveal anything about who gets their money. This explanation can be made for the **Economy** example, too.[11] In other cases, like **Student Housing** or **Ice-Skating**, the conflict doesn't seem to destroy or harm anything over and above the conflicting part itself. We have some information that speaks in favor of Brigt Rykkje being Norwegian and we have some other information that speaks in favor of Brigt Rykkje being Dutch. However, the different kinds and sources of information are independent of each other and are not destroyed by the conflict. In all of the cases where the conflict is harmful to the general argument, this is so because the *credibility* of the sources of information or the *authority* of the experts is undermined by the conflict.[12] It is not clear, however, how this explanation succeeds to explain the different intuitions about **Mary in Canada** and **Carol in the US**, or **Wavelength Color** and **Primary Color**. The conflict involved is exactly of the same form, and thus, it is not clear why the conflict is harmful for one floating conclusion but not for the other.

4 A possible solution: A default framework for floating conclusions

In the last section, we found that none of the presented hypotheses is suited to explain the intuitions about *all* examples. Still, we are positive that the two hypotheses "Possible Compromise" and "Harmfulness of the Conflict" combined manage to describe what is at the heart of the matter for the different examples. For example, "Possible Compromise" nicely explains the different judgments for **Mary in Canada** and **Carol in the US** and for **Wavelength Color** and **Primary Color** by referring to the compromising proposition.

We propose a solution that manages to combine different hypotheses. The basic idea is that a floating conclusion should be accepted by default. We should prima facie believe them. Then, there are different reasons to deviate from the default and to reject a floating conclusion. One such reason is explained by the "Possible Compromise" thesis. If there is a compromise between the conflicting propositions (and from this compromising proposition the floating conclusion doesn't follow) then the floating conclusion is to be rejected.[13] Another reason to deviate from the default and to reject the floating conclusion is described

[11] It would be interesting to see if the situation changes when the two conflicting propositions that seem to 'cancel each other out' are not equally strong.

[12] This can be seen even clearer when modeling examples like **Yacht** or **Murderer** in a different way. For example, one can take e.g. "Witness A says that p, hence p" to be not a defeasible argument but rather a justification through testimony or utterance for p. The argument as such then starts from the two premises "Peter killed the victim with a knife" and "Peter killed the victim with a gun" (which are both justified by some testimony). Then one could argue that both arguments (from K to V and from G to V) are in fact even *undermined* (see [8] for this terminology) since the premises of the arguments are attacked by the conflicting testimonies. Many thanks to Stipe Pandzic for this remark.

[13] It can be noted that the vagueness thesis describes a special case of a possible compromise.

in the “Harmfulness of the conflict” thesis. If a conflict is not only harmful to the conflicting propositions but undermines the credibility or authority of the sources of information entirely, then the floating conclusion is to be rejected. The basic idea behind this can already be found in [5, p. 189]: “we might suppose that, although floating conclusions are in general acceptable, there are structural features present in situations such as the yacht example, but not yet captured in our formal representations of these examples that block these conclusions.”

In all cases of floating conclusions, the conflicting propositions are contrary to each other. Although they cannot both be true at the same time, they can both be false at the same time. That is, in addition to the possibility that one proposition is true and the other false (or vice versa), there is a third possibility: both propositions are false. The two reasons to deviate from the default describe both one version of (or reasons for) this third possibility. Either we reject both propositions because the credibility of their justification has been undermined or because there is a third proposition as a compromise available.[14] Logically, one could capture this by saying that both conflicting propositions p and q that lead to a floating conclusion are false, i.e., $\neg p \wedge \neg q$ or $\neg(p \vee q)$. The other (default) situation in which we should accept the floating conclusion could then be captured by the *exclusive disjunction* $p \veebar q$ of the two propositions being true.[15] With this manifold solution, we think that we manage best to precisely describe what is going on in the different examples and hit the heart of the matter, describing the underlying patterns of the intuitions. In the case of **Primary Color**, **Nixon**, or **Carol in the US** it is really the plausibility of the compromise between the conflicting propositions (either that the cup is orange, that Nixon is politically in the middle, or that Carol lives ‘in between’) that makes us reject the floating conclusion intuitively. In the cases of **Yacht** and **Murderer** or **Economy**[16] the reason why we intuitively reject the floating conclusion is that we do not trust any line of argument anymore as the credibility of the sources got destroyed. For example, in the **Murderer** case, the credibility of the testimonies is destroyed by their disagreeing about the weapon. Moreover, we do not want to claim that these two reasons: compromise and harmfulness of the conflict are the only reasons for deviating from the default of accepting the floating conclusion. Plausibly, there will be other reasons. This is not a problem for our theory, though, as this default-based framework can easily be extended with multiple more reasons to deviate.

[14] In this sense, the “hidden default” thesis can also be incorporated into the framework. The two explanations ”possible compromise” and “harmfulness of the conflict” describe different reasons why in some examples a default still seems to be missing or hidden.

[15] Thanks to an anonymous reviewer for pointing this out.

[16] In fact, **Economy** can be explained both by referring to a possible compromise *and* by the harmfulness of the conflict for the credibility of the sources. Thus, this example shows that there can be even more than one reason to deviate from the default of accepting floating conclusions.

5 Conclusion and Outlook

In this paper, we investigated the phenomenon of floating conclusions. The question about the acceptability of floating conclusions can be reformulated as the question of whether we should accept at least one line of reasoning among a set of conflicting lines of reasoning. We presented an overview of different examples of floating conclusions from the literature and extended the list with new examples. We examined different hypotheses that aim to explain our non-uniform intuitions about whether floating conclusions should be accepted or not and tested them via our examples. We argued that no hypothesis succeeds in explaining our intuitions concerning *all* the presented examples. Instead, we presented an overarching explanation for the acceptability of floating conclusions. The explanation starts with the basic idea that floating conclusions ought to be accepted *by default.* The framework then allows several reasons to deviate from the default and to reject the floating conclusion. These reasons come into play when there seems to be a third alternative besides the two conflicting propositions. We presented two possible reasons why this alternative arises. If there is a compromise between the conflicting propositions from which the floating conclusion does not follow or if the conflict is harmful to the sources of information, one can deviate from the default and reject the floating conclusion. We saw that these two reasons nicely cover and explain all the examples investigated in this paper. We thereby manage to describe the underlying pattern of our intuitions regarding the floating conclusions. Still, the framework is open for new, additional reasons that will come along with new examples when the matter is investigated more.

As we already mentioned in the introduction, intuitions alone do not always help us decide about the different examples. This is visualized nicely in the following example from practical reasoning. Imagine there was a robbery where jewelry was stolen. Later, the police stop a man in a car and find the stolen jewelry. The police have reason to believe that the occupant stole the jewelry. However, the man claims to have bought the jewelry. Both activities (stealing and the so-called 'Hehlerei'/'heling,' i.e., the purchase of stolen goods) are punishable in the Netherlands as well as in Germany. The German legal system allows the suspect to be convicted for the crime with the lesser penalty since it is clear that he committed one of the two crimes. The Dutch legal system, on the other hand, cannot convict the suspect unless there is evidence that clearly shows which of the crimes was committed.[17] The acceptance of the practical floating conclusion (the suspect is punishable) here does not depend on intuitions but on the legal system, one is referring to. The dependency on context and on stakes can also be nicely visualized by our presented examples. While the conflict destroys the credibility of the witnesses in **Yacht** or **Murderer**, the conflict does not seem to destroy the credibility of Anna and Ben in the color examples. In these contexts, where they are simply telling us the color of a cup, we have no reason to be suspicious because the context offers

[17] According to: Hans Nijboer, personal communication, 2007

us no reason why they should lie about the color of the cup. [18] Since whether or not we want to choose the third alternative, deviate from the default, and reject the floating conclusion seems to depend heavily on the stakes and on the context, we consider it a very difficult challenge to represent the appropriate handling of floating conclusions in a formal logical system. Moreover, these examples suggest that there might be a difference between purely theoretical, epistemological reasoning, and practical reasoning. As the intuitions would also become more comparable when actions are involved, further research on the influence of practical reasoning for floating conclusions seems very fruitful.

References

[1] Brandom, R., "Articulating Reasons: An introduction to inferentialism," Harvard University Press, Cambrdige, Massachusetts, 2000.

[2] Broersen, J., *Rethinking the boid with an eye on making it more responsible*, manuscript (unpublished).

[3] Ginsberg, M. L., "Essentials of Artificial Intelligence," Morgan Kaufmann Publishers In, San Francisco, 1993.

[4] Horty, J. F., *Skepticism and floating conclusions*, Artificial Intelligence **135** (2002), pp. 55–72.

[5] Horty, J. F., "Reasons as Defaults," Oxford University Press, Oxford, 2011.

[6] Horty, J. F., R. H. Thomason and D. S. Touretzky, *A skeptical theory of inheritance in nonmonotonic semantic networks*, Artificial intelligence **42** (1990), pp. 311–348.

[7] Makinson, D. and K. Schlechta, *Floating conclusions and zombie paths: two deep difficulties in the "directly skeptical" approach to defeasible inheritance nets*, Artificial intelligence **48** (1991), pp. 199–209.

[8] Pollock, J. L., "Cognitive carpentry: A blueprint for how to build a person," MIT Press, Cambridge, Mass, 1995.

[9] Prakken, H., *Intuitions and the modelling of defeasible reasoning: some case studies*, in: *Proceedings of the 9th International Workshop on Non-Monotonic Reasoning. Toulouse, France*, 2002.

[10] Reiter, R., *A logic for default reasoning*, Artificial intelligence **13** (1980), pp. 81–132.

[11] Sellars, W., *Inference and meaning*, Mind **62** (1953), pp. 313–338.

[12] Veltman, F., "Logics for conditionals," Thesis (1985).

[18] Thanks to Joris Graff for coming up with this point.

Strong Permission Bundled: First Steps

Zilu Wang Yanjing Wang

{ziluwang19, y.wang}@pku.edu.cn
Department of Philosophy
Peking University

Abstract

In this paper, we introduce a novel framework for deontic logic of strong permission that accommodates free choice. Our approach treats permission as a *bundled modality*, which combines a universal quantifier with a possibility modality such that an action type α is permitted if and only if *every* token of α can be executed in *some* deontically ideal world. Our formalization of action tokens and their types is inspired by the BHK-style interpretation for intuitionistic logic. We axiomatize the logics of strong permission under various conditions. Beyond satisfying desirable logical properties found in the literature, our framework also predicts interesting new phenomena related to permission and distribution laws that align with our linguistic intuition.

Keywords: deontic logic, bundled modality, strong permission, free choice, BHK interpretation, first-order modal logic

1 Introduction

Modalities are often more than what they appear to be. An innocent-looking modality in natural language may have hidden inner logical structures that can cause its behavior to diverge from normal modal logic, resulting in a range of puzzles. For example, considering know-how as a $\Box$ modality, one can know how to achieve α ($\Box\alpha$) and how to achieve β ($\Box\beta$), without knowing how to achieve $\alpha \wedge \beta$ simultaneously ($\Box(\alpha \wedge \beta)$). Thus the conjunction aggregation in normal modal logic is intuitively invalid. Technically, such non-normality can be accommodated by more general semantics (cf. e.g., [9]), but it has a deeper semantic root that the know-how modality can be understood as a *bundle* of an existential quantifier and a know-that modality $\exists x\mathbf{K}$, i.e., knowing how to achieve α can be interpreted as *there is* a plan such that one *knows that* it is executable and will guarantee α (cf. e.g., [23]). The interaction between $\exists$ and $\mathbf{K}$ reflects the *de re* nature of knowledge-how and can also account for its "ambiguous" logical behavior ensembles both $\Box$ and $\Diamond$ to some extent.[1]

Such cases may find resonance in deontic logic, where logical puzzles abound. One of the most discussed puzzles in deontic logic is the puzzle of

[1] The swapped version $\mathbf{K}\exists x$ represents *de dicto* knowledge, e.g., knowing that *there is* a proof for some theorem, which does not imply knowing how to prove the theorem.

free choice permission (FCP) [14], wherein the permission to do either α *or* β intuitively results in the permission of α *and* the permission of β in various contexts. This view contrasts with the notion that permission is a $\Diamond$-like modality in standard deontic logic (SDL). Furthermore, incorporating FCP as an axiom on top of SDL yields unacceptable consequences. This raises the question of whether there is an unexplored logical inner structure of permission that can account for FCP and other related puzzles.

We think the answer is positive. In particular, there may be a hidden bundle of a quantifier and a modality behind the modality of permission, as in the case of know-how. Actually, this is *not* a new idea in deontic logic.

First of all, quantifiers can be introduced to deontic logic over the set of agents to whom the norms are applied. However, a more significant role played by quantifiers emerges when we distinguish between *action tokens* and *action types.* This distinction was already observed by von Wright in his seminal work that pioneered the field of deontic logic [24]. Action tokens can be considered individual acts of certain action types, with deontic modalities primarily applied to the latter. The distinction between types and tokens has proven fruitful in addressing puzzles like FCP, such as the Boolean-algebra-based approach initiated by Segerberg [19] and its more sophisticated generalizations based on (il)legal sets [8,20] (see also [12,6,4,10] for comprehensive reviews of the literature regarding approaches to FCP). Like many other puzzles in deontic logic, FCP cannot be solved in isolation. It remains to be seen whether a semantic approach can lead to a logic satisfying most, if not all, of the intuitive constraints discussed in the literature. [2] This paper presents the first steps of our attempt toward this goal.

Our approach is inspired by Hintikka, who explored various combinations of quantifiers and modalities to address the deficiencies of SDL in the early days of deontic logic [13]. Given the distinction between action types and tokens, we can quantify the tokens in defining the semantics for permission in combination with the modality. In particular, Hintikka informally proposed capturing a notion of permission with $\forall x \Diamond \alpha$, i.e., a (strong) permission of α asks for each token of type α to be executed in some deontically ideal world w.r.t. the current world. For instance, if you are allowed to take one day off next week, it means that you can take *any* single day off next week. If you are only allowed to take Monday off, then the previous permission would be considered false, i.e., allowing merely one token does not justify the strong permission about the action type. Universally quantified interpretations of permission can also be found in other approaches, such as [5] in terms of “open reading of permission”, and [18,7] in the tradition of dynamic deontic logics. [3]

[2] There are “hybrid” approaches that combine semantics and pragmatic features, satisfying most constraints (cf., e.g. [2]).

[3] Some may view the use of universal quantifier to be too strong as it may involve *every* token no matter how strange, cf., e.g., [6]. Note that as the interpretation of quantifiers in other applications, it is always about a set of *relevant* acts given in the context (technically a model). We leave the philosophical discussions on this to the full version of the paper.

However, Hintikka's original proposal was mostly forgotten in the literature after receiving early criticisms regarding the use of quantifiers (cf. e.g., [16]).

We believe that Hintikka's $\forall\Diamond$-schema for permission has the potential to provide a satisfactory semantics for strong permission. However, one of the missing pieces is an alternative formal treatment of action types and tokens, which would differ from those found in the literature, such as the ones based on Boolean algebra. Our inspiration comes from the Brouwer-Heyting-Kolmogorov (BHK) interpretation of intuitionistic logic, which has an interesting connection with the know-how bundle [22]. Although this interpretation is commonly known as the *proof* interpretation, a more general view advocated by Kolmogorov treats each intuitionistic formula as a *type* of problem, interpreting it as the set of all its *solutions* [15,17]. From this broader perspective, formulas are not necessarily propositions with truth values; rather, they represent types of specific entities. We draw upon this parallel and provide a semantics for strong permissions, utilizing the ideas in [22] to formalize the $\forall\Diamond$ bundle instead of the $\exists\Box$ bundle for the know-how operator found in [22].

In this initial attempt, we constrain our language and focus solely on the concept of strong permission, which is taken to be descriptive rather than declarative [12]. Technically, our contribution encompasses a new formal semantics, along with a series of complete axiomatizations for the logic of strong permission under various settings. These complete axiomatizations with intuitive axioms largely assure us that there are no unexpected consequences arising from our semantics. Moreover, as we will explain below, our approach uncovers intriguing natural language phenomena related to strong permission, which might have remained unnoticed without such a formal framework.

As a preview, we first list the logical features of our framework:

Valid in our framework			
`FC`	$\mathbf{P}(\alpha \vee \beta) \leftrightarrow (\mathbf{P}\alpha \wedge \mathbf{P}\beta)$	`CD`	$\mathbf{P}(\alpha \wedge (\beta \vee \gamma)) \leftrightarrow \mathbf{P}((\alpha \wedge \beta) \vee (\alpha \wedge \gamma))$
`CE`	$\mathbf{P}(\alpha \wedge \beta) \to (\mathbf{P}\alpha \wedge \mathbf{P}\beta)$	`DCl`	$\mathbf{P}((\alpha \vee \beta) \wedge (\alpha \vee \gamma)) \to \mathbf{P}(\alpha \vee (\beta \wedge \gamma))$
Invalid in our framework			
`CA`	$(\mathbf{P}\alpha \wedge \mathbf{P}\beta) \to \mathbf{P}(\alpha \wedge \beta)$	`DCr`	$\mathbf{P}(\alpha \vee (\beta \wedge \gamma)) \to \mathbf{P}((\alpha \vee \beta) \wedge (\alpha \vee \gamma))$
`RP`	$\mathbf{P}\alpha \to \mathbf{P}(\alpha \vee \beta)$	`EX`	$\mathbf{P}\alpha \to \mathbf{P}(\alpha \wedge \alpha)$
`MN`	$\vdash \alpha \to \beta \Rightarrow \vdash \mathbf{P}\alpha \to \mathbf{P}\beta$	`RE`	$\vdash \alpha \leftrightarrow \beta \Rightarrow \vdash \mathbf{P}\alpha \leftrightarrow \mathbf{P}\beta$

`FC` is the (two-way) free choice principle; `CE` is the conjunction exploitation. Three of the four (one-way) distribution laws are valid as captured by `CD` and `DCl`, leaving `DCr` invalid. The invalidity of `DCr` is intuitive: e.g., imagine you are given a coupon that allows you to take a hamburger *or* a menu of French fries *and* salad, this does not mean you can take a hamburger *or* fries, *and* a hamburger *or* salad. Note that the later permission intuitively allows you to take a hamburger and salad, which is not allowed by the premise (the coupon). The conjunction aggregation `CA` is intuitively invalid as you may not be allowed to do α and β together, given the permission of each. Moreover, `RP` captures the invalid Ross's paradox for permission. `EX` is the duplication of permitted actions, whose invalidity may raise some eyebrows as it looks so innocent.

However, the case is more subtle than it may appear. According to our $\forall\Diamond$ semantics for $\mathbf{P}$, allowing α does not mean you can do α twice using any combination of the tokens. More importantly, adding it as an axiom leads to a very counterintuitive consequence that $\mathbf{P}(\alpha \vee \beta) \to \mathbf{P}(\alpha \wedge \beta)$, given the widely accepted `CD` and `FC`:

$$\begin{array}{lll}
 & \mathbf{P}(\alpha \vee \beta) & \\
\Longrightarrow & \mathbf{P}((\alpha \vee \beta) \wedge (\alpha \vee \beta)) & (\mathtt{EX}) \\
\Longleftrightarrow & \mathbf{P}(((\alpha \vee \beta) \wedge \alpha) \vee ((\alpha \vee \beta) \wedge \beta)) & (\mathtt{CD}) \\
\Longleftrightarrow & \mathbf{P}((\alpha \vee \beta) \wedge \alpha) \wedge \mathbf{P}((\alpha \vee \beta) \wedge \beta) & (\mathtt{FC}) \\
\Longleftrightarrow & \mathbf{P}((\alpha \wedge \alpha) \vee (\alpha \wedge \beta)) \wedge \mathbf{P}((\beta \wedge \alpha) \vee (\beta \wedge \beta)) & (\mathtt{CD}, \text{commutativity}) \\
\Longleftrightarrow & \mathbf{P}(\alpha \wedge \alpha) \wedge \mathbf{P}(\alpha \wedge \beta) \wedge \mathbf{P}(\beta \wedge \alpha) \wedge \mathbf{P}(\beta \wedge \beta) & (\mathtt{FC}) \\
\Longrightarrow & \mathbf{P}(\alpha \wedge \beta) & (\mathtt{TAUT})
\end{array}$$

Nevertheless, a weaker version of `EX` can be made valid by imposing some further constraints in our framework. Finally, it is important to note that the rules of monotonicity (`MN`) and replacement of equivalents (`RE`), which often cause counterintuitive consequences [12,11], are *not* valid, as demonstrated by the fact that classical tautologies are valid in our setting, but `EX` and `DCr` are not.

To the best of our knowledge, these (in)validities distinguish our logic from existing approaches that admit FCP. For instance, `CA` and `DCr` are valid in Boolean-algebra-based approaches, such as [8,20]; `CD` is invalid in [6] due to the emptiness condition for disjunction; `DCr` is also valid in the hybrid approach based on BSML [2]; and `CE` is not valid in [21]. We believe that an effective semantics should not merely serve as a technical tool to "fit" the known behaviors of a modality, but also correctly predict intriguing new phenomena that align with its use in natural language.

The rest of the paper is structured as follows. In Section 2, we lay out the fundamentals of our framework and introduce the proof systems $\mathbb{DLSP}$ and $\mathbb{DLSP}^s$. Section 3 presents the completeness proofs. In Section 4, we discuss the generalizations of our models and provide the corresponding axiomatization results. Finally, we conclude the paper and outline future directions in Section 5.

2 Language, Semantics and Proof System

2.1 Language and Semantics

As mentioned in the introduction, we use a propositional language to specify the action types. As the initial step, we only include constructors of $\vee$ and $\wedge$, and we will come back to other constructors at the end of the paper.

Definition 1 (Action Type AT) *Given a countable set P of propositional letters, the language of action types ($\mathbf{AT}_P$) is defined as follows:*

$$\alpha ::= p \mid (\alpha \wedge \alpha) \mid (\alpha \vee \alpha)$$

where $p \in P$.

Propositional letters are intended to represent *atomic* action types; $\alpha \wedge \beta$ represents a joint action type of doing both α and β (though not necessarily at the same time); $\alpha \vee \beta$ is a type of doing α or β. Given P (and thus $\mathbf{AT}_P$), we define the language of our deontic logic for strong permission.

Definition 2 (Language DLSP) *Given P, the language of deontic logic for strong permission ($\mathbf{DLSP}_P$) is defined as follows:*

$$\varphi ::= \bot \mid p \mid (\varphi \wedge \varphi) \mid (\varphi \vee \varphi) \mid (\varphi \to \varphi) \mid \neg\varphi \mid \mathbf{P}\alpha,$$

where $p \in P$ and $\alpha \in \mathbf{AT}_P$.

$\mathbf{P}\alpha$ says that action type α *is (strongly) permitted*. As we will see later, the connectives outside $\mathbf{P}$ are classical. In the following we fixed a countably infinite P, and write $\mathbf{DLSP}_P$ and $\mathbf{AT}_P$ simply as $\mathbf{DLSP}$ and $\mathbf{AT}$.

Before introducing the model, we define the action tokens of each type inspired by the BHK interpretation (cf., e.g., [17]).

Definition 3 (Action Token Space) *Given P and a non-empty set I of atomic action tokens such that $I \cap \{0, 1\} = \emptyset$, an action (token) space S based on I is a function over $\mathbf{AT}$ satisfying the following constraints:*

(i) *$S(p) \neq \emptyset \subseteq I$ for any $p \in P$;*

(ii) *$S(\alpha \wedge \beta) = S(\alpha) \times S(\beta)$;*

(iii) *$S(\alpha \vee \beta) = (S(\alpha) \times \{0\}) \cup (S(\beta) \times \{1\})$.*

S gives the (non-empty) set of possible tokens of each composite type based on tokens for the atomic types. A token of a disjunctive type is a pair of a token of one of the disjunct types and a marker 0 or 1 indicating the left or right disjunct that it belongs to. The tokens of a conjunctive type are exactly those pairs of tokens from both conjunct types respectively.

As an interesting special case, a *singleton* action space reflects the setting where atomic types and tokens are not distinguished.

Definition 4 (Singleton Action Token Space) *A singleton action token space S is an action space such that for any $p \in P$, $S(p)$ is a singleton.*

Now we can define the deontic model where tokens in the space are executed or not on each possible idealization of the current world.

Definition 5 (Deontic Model) *A deontic model $\mathcal{M}$ for* **DLSP** *is a tuple (S, W, R, A) where S is an action space based on some non-empty I, W is a non-empty set of possible worlds, $R \subseteq W \times W$, and A is a binary function over $\mathbf{AT} \times W$ such that for any $p \in P$, $\alpha, \beta \in \mathbf{AT}$ and $w \in W$:*

- *$A(p, w) \subseteq S(p)$;*
- *$A(\alpha \wedge \beta, w) = A(\alpha, w) \times A(\beta, w)$;*
- *$A(\alpha \vee \beta, w) = (A(\alpha, w) \times \{0\}) \cup (A(\beta, w) \times \{1\})$;*

A pointed deontic model is an ordered pair $(\mathcal{M}, w)$ where w is in $\mathcal{M}$. A singleton *deontic model is a model based on a singleton action space.*

Intuitively, wRv means the world v is a deontic idealization for w.[4] The function A gives each world its *executed* (or *realized*) action tokens compositionally for each type within its action space, e.g., if tokens a of type α and b of type β are both executed on world w, then (a, b) of type $\alpha \wedge \beta$ is executed on w.

Definition 6 (Semantics) *For any $\varphi \in$ **DLSP** and any pointed deontic model $\mathcal{M}, w$ where $\mathcal{M} = (S, W, R, A)$, the satisfaction relation is defined as:*

$$\boxed{\begin{array}{lll}
\mathcal{M}, w \nvDash \bot & & \\
\mathcal{M}, w \vDash p & \iff & A(p, w) \neq \emptyset \\
\mathcal{M}, w \vDash (\varphi \wedge \psi) & \iff & \mathcal{M}, w \vDash \varphi \textit{ and } \mathcal{M}, w \vDash \\
\mathcal{M}, w \vDash (\varphi \vee \psi) & \iff & \mathcal{M}, w \vDash \varphi \textit{ or } \mathcal{M}, w \vDash \\
\mathcal{M}, w \vDash (\varphi \to \psi) & \iff & \mathcal{M}, w \nvDash \varphi \textit{ or } \mathcal{M}, w \vDash \\
\mathcal{M}, w \vDash \neg\varphi & \iff & \mathcal{M}, w \nvDash \varphi \\
\mathcal{M}, w \vDash \mathbf{P}\alpha & \iff & \textit{for any } a \in S(\alpha), \textit{ there is a } v \textit{ s.t.} wRv \textit{ and } a \in A(\alpha, v)
\end{array}}$$

The notion of semantic consequence $\Gamma \vDash \varphi$ are defined as usual. We use $\vDash_s$ to denote semantic consequence w.r.t. singleton deontic models. We say φ is valid (s-valid) if $\vDash \varphi$ ($\vDash_s \varphi$).

Note that the boolean connectives outside **P** are indeed interpreted classically. On the other hand, the semantics of **P** involves a bundle $\forall\Diamond$ that we mentioned, thus making $\mathbf{P}\alpha$ true iff *each* token of α can be executed on *some* deontically ideal world w.r.t. the current world. Based on the semantics of p and the definition of A for the $\wedge$ and $\vee$ types, we can show that:

Proposition 2.1 *For any $\alpha \in$ **AT**, any pointed deontic model $\mathcal{M}, w$, $\mathcal{M}, w \vDash \mathbf{P}\alpha$ implies there is v s.t., wRv and $\mathcal{M}, v \vDash \alpha$.*

However, the converse is not true, e.g., $\mathcal{M}, v \vDash p$ and wRv do not imply $\mathcal{M}, w \vDash \mathbf{P}p$, for the latter requires every token of p to be realized. Thus **P** is not reducible to a standard $\Diamond$. Note that, in general, the truth value of α in the actual world has no connection with the truth value of $\mathbf{P}\alpha$, e.g., $\neg\alpha \wedge \mathbf{P}\alpha$ and $\alpha \wedge \neg\mathbf{P}\alpha$ are both satisfiable. To familiarize the readers with the semantics, let us check the validity of a few formulas such as `FC`.

Proposition 2.2 *For any $\alpha, \beta \in$ **AT**, the following schemas are valid:*

$$\texttt{FC} \quad \mathbf{P}(\alpha \vee \beta) \leftrightarrow (\mathbf{P}\alpha \wedge \mathbf{P}\beta) \qquad \texttt{CE} \quad \mathbf{P}(\alpha \wedge \beta) \to (\mathbf{P}\alpha \wedge \mathbf{P}\beta)$$

Proof. We only present the proof for the validity of `FC` here. Let $\mathcal{M} = (S, W, R, A)$ be a deontic model and $w \in W$ be arbitrary.

$\Rightarrow$: Assume that $\mathcal{M}, w \vDash \mathbf{P}(\alpha \vee \beta)$. We only show $\mathcal{M}, w \vDash \mathbf{P}\alpha$ for the case of $\mathbf{P}\beta$ is symmetric. Given an arbitrary $a \in S(\alpha)$, it follows that $(a, 0) \in S(\alpha \vee \beta)$. Since $\mathcal{M}, w \vDash \mathbf{P}(\alpha \vee \beta)$, there is a $v \in W$ such that wRv and $(a, 0) \in A(\alpha \vee \beta, v)$. By the definition of A in $\mathcal{M}$, $a \in A(\alpha, v)$. Since the selection of a is arbitrary, we have $\mathcal{M}, w \vDash \mathbf{P}\alpha$.

[4] Imposing seriality in the model will result in the same logic. See later discussion.

$\Leftarrow$: Assume that $\mathcal{M}, w \vDash (\mathbf{P}\alpha \wedge \mathbf{P}\beta)$. Given an arbitrary $x \in S(\alpha \vee \beta)$, it follows that either $x = (a, 0)$ for some $a \in S(\alpha)$ or $x = (b, 1)$ for some $b \in S(\beta)$ by the definition of S. Assume without loss of generality that $x = (a, 0)$. Since $\mathcal{M}, w \vDash \mathbf{P}\alpha$, there is a $v \in W$ such that wRv and $a \in A(\alpha, v)$. By the definition of A in $\mathcal{M}$, $(a, 0) \in A(\alpha \vee \beta, v)$. □

Given the above proposition, it is clear that $\mathbf{P}(\alpha \wedge \beta) \to \mathbf{P}(\alpha \vee \beta)$. However the converse is not valid as the following proposition implies.

Proposition 2.3 *For any $\alpha, \beta \in \mathbf{AT}$, the following formulas are **not** valid:*

CA $(\mathbf{P}\alpha \wedge \mathbf{P}\beta) \to \mathbf{P}(\alpha \wedge \beta)$ *EX* $\mathbf{P}\alpha \to \mathbf{P}(\alpha \wedge \alpha)$ *RP* $\mathbf{P}\alpha \to \mathbf{P}(\alpha \vee \beta)$

Proof. For CA, note that it is possible that every token a of α and every token b of β are executable respectively in accessible ideal worlds but some combination $(a, b) \in S(\alpha \wedge \beta)$ is not executable in *any* accessible world.

For EX, let $S(p) = \{a, b\}$ then $S(p \wedge p) = \{(a, a), (b, b), (a, b), (b, a)\}$. It is easy to define a model $\mathcal{M}, w$ such that $\mathcal{M}, w \vDash \mathbf{P}p \wedge \neg \mathbf{P}(p \wedge p)$ with no accessible ideal world witnessing both tokens; thus (a, b) is not executable anywhere.

The invalidity of RP is clear given that $\mathbf{P}(\alpha \vee \beta) \leftrightarrow (\mathbf{P}\alpha \wedge \mathbf{P}\beta)$ is valid. □

The invalidity of CA prevents $\mathbf{P}(\alpha \wedge \beta)$ from being equivalent to $\mathbf{P}(\alpha \vee \beta)$. Given the proof for EX, readers may wonder what would happen if the atomic type p has at most one token. We will address this later in Section 2.2. The invalidity of RP avoids Ross' paradox in the context of strong permission. Further, note that under our semantics, $\vDash \alpha \leftrightarrow (\alpha \wedge \alpha)$ and $\vDash \alpha \to (\alpha \vee \beta)$ hold for any $\alpha, \beta \in \mathbf{AT}$. Therefore, it is not hard to see that:

Proposition 2.4 *The following rules are invalid:*

(i) *Monotonicity (MN): for any $\alpha, \beta \in \mathbf{AT}$, if $\vdash \alpha \to \beta$, then $\vdash \mathbf{P}\alpha \to \mathbf{P}\beta$.*

(ii) *Replacement of equivalents (RE): for any $\alpha, \beta \in \mathbf{AT}$, if $\vdash \alpha \leftrightarrow \beta$, then $\vdash \mathbf{P}\alpha \leftrightarrow \mathbf{P}\beta$.*

The commutativity and associativity laws are valid within the scope of $\mathbf{P}$, and we omit the proofs that are routine checks by definition.

Proposition 2.5 *For any $\alpha, \beta \in \mathbf{AT}$, the following formulas are valid:*

$$\begin{array}{ll} \mathbf{P}(\alpha \wedge \beta) \leftrightarrow \mathbf{P}(\beta \wedge \alpha) & \mathbf{P}((\alpha \wedge \beta) \wedge \gamma) \leftrightarrow \mathbf{P}(\alpha \wedge (\beta \wedge \gamma)) \\ \mathbf{P}(\alpha \vee \beta) \leftrightarrow \mathbf{P}(\beta \vee \alpha) & \mathbf{P}((\alpha \vee \beta) \vee \gamma) \leftrightarrow \mathbf{P}(\alpha \vee (\beta \vee \gamma)) \end{array}$$

Concerning the distributivity within $\mathbf{P}$, the situation is more subtle. As we mentioned in the introduction, one of the four (one-way) laws is invalid.

Proposition 2.6 *For any $\alpha, \beta, \gamma \in \mathbf{AT}$, we have:*

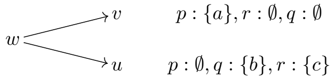
CD: $\vDash \mathbf{P}(\alpha \wedge (\beta \vee \gamma)) \leftrightarrow \mathbf{P}((\alpha \wedge \beta) \vee (\alpha \wedge \gamma))$;

DCr: $\nvDash \mathbf{P}(\alpha \vee (\beta \wedge \gamma)) \to \mathbf{P}((\alpha \vee \beta) \wedge (\alpha \vee \gamma))$;

DCl: $\vDash \mathbf{P}((\alpha \vee \beta) \wedge (\alpha \vee \gamma)) \to \mathbf{P}(\alpha \vee (\beta \wedge \gamma))$.

Proof. We only show the invalidity of `DCr` and the validity of `DCl`. For `DCr`, consider the countermodel illustrated below, where the rightmost part demonstrates the definition of A on u, v, e.g., $A(p, v) = \{a\}$ and $A(q, u) = \{b\}$.

$S(p) = \{a\}, S(q) = \{b\}, S(r) = \{c\}$


w
v
u
$p : \{a\}, r : \emptyset, q : \emptyset$
$p : \emptyset, q : \{b\}, r : \{c\}$

$S(p \vee (q \wedge r))$ contains $(a, 0)$ and $((b, c), 1)$ only, which are executable on v and u respectively, thus $\mathbf{P}(p \vee (q \wedge r))$ is true on w. However, the token $((a, 0), (c, 1))$ in $S((p \vee q) \wedge (p \vee r))$ is not executable on u nor v, thus $\mathbf{P}((p \vee q) \wedge (p \vee r))$ is false on w. Note that this model is also a singleton model so `DCr` is not s-valid.

For `DCl`, let $\mathcal{M} = (S, W, R, A)$ be a deontic model and $w \in W$ be arbitrary. Assume that $\mathcal{M}, w \vDash \mathbf{P}((\alpha \wedge \beta) \vee (\alpha \wedge \gamma))$. Let $x \in S(\alpha \wedge (\beta \vee \gamma))$ be arbitrary. Thus $x = (a, (b, 0))$ for some $a \in S(\alpha)$ and $b \in S(\beta)$, or $x = (a', (c, 1))$ for some $a' \in S(\alpha)$ and $c \in S(\gamma)$. Assume without loss of generality that $x = (a, (b, 0))$. Since $((a, b), 0) \in S((\alpha \wedge \beta) \vee (\alpha \wedge \gamma))$, there is a $v \in W$ such that wRv and $((a, b), 0) \in A((\alpha \wedge \beta) \vee (\alpha \wedge \gamma), v)$. By definition, $a \in A(\alpha, v)$ and $b \in A(\beta, v)$. Therefore, $(a, (b, 0)) \in A(\alpha \wedge (\beta \vee \gamma), v)$. □

Finally, let us consider the singleton action space and the corresponding singleton models, where a variant of `EX` in Proposition 2.3 is s-valid. In the following, we use $m_i \cdot p_i$ to abbreviate the conjunction of m_i copies of p_i.[5]

Proposition 2.7 *The following formula (denoted by `EXP`) is valid with respect to the class of singleton deontic models:*

$$\vDash_s \mathbf{P}(p_1 \wedge ... \wedge p_k) \to \mathbf{P}(m_1 \cdot p_1 \wedge ... \wedge m_k \cdot p_k),$$

where $p_1, ...p_k \in P$ are distinct, and $k, m_i \in \mathbb{N}_{>0}$ for any $1 \leq i \leq k$.

Proof. Let $\mathcal{M} = (S, W, R, A)$ be an arbitrary singleton deontic model and $w \in W$. Since for any $1 \leq i \leq k$, $S(p_i)$ is a singleton, $S(p_1 \wedge ... \wedge p_k) = \Pi^k_{i=1} S(p_i)$ is also a singleton. Assume that $\mathcal{M}, w \vDash \mathbf{P}(p_1 \wedge ... \wedge p_k)$. Thus there is a $v \in W$ such that wRv and $A((p_1 \wedge ... \wedge p_k), v) = \Pi^k_{i=1} S(q_i)$. Hence, by definition, for any $1 \leq i \leq k$, $A(p_i, v) = S(p_i)$. So, $A((m_1 \cdot p_1 \wedge ... \wedge m_k \cdot p_k), v) = \Pi^{m_1}_{i=1} S(p_1) \times ... \times \Pi^{m_k}_{i=1} S(p_k) = S(m_1 \cdot p_1 \wedge ... \wedge m_k \cdot p_k)$. Therefore, $\mathcal{M}, w \vDash \mathbf{P}(m_1 \cdot p_1 \wedge ... \wedge m_k \cdot p_k)$.□

Note that the converse holds by the validity of `CE`. Over singleton models, `EX` is still *not* valid, e.g., $\nvDash_s \mathbf{P}(p \vee q) \to \mathbf{P}((p \vee q) \wedge (p \vee q))$ since $S(p \vee q)$ is not a singleton.

2.2 Proof Systems and Normal Form

Here we introduce the following two proof systems.

System $\mathbb{DLSP}$

[5] Notations like $\mathbf{P}(p_1 \wedge ... \wedge p_k)$ are justified because action conjunctions within $\mathbf{P}$ are commutative and associative under the above semantics.

Axioms	
(TAUT)	Propositional Tautologies
(FC)	$\mathbf{P}(\alpha \vee \beta) \leftrightarrow (\mathbf{P}\alpha \wedge \mathbf{P}\beta)$
(CE)	$\mathbf{P}(\alpha \wedge \beta) \rightarrow (\mathbf{P}\alpha \wedge \mathbf{P}\beta)$
($\mathtt{COM}_\wedge$)	$\mathbf{P}(\alpha \wedge \beta) \leftrightarrow \mathbf{P}(\beta \wedge \alpha)$
($\mathtt{ASSO}_\wedge$)	$\mathbf{P}((\alpha \wedge \beta) \wedge \gamma) \leftrightarrow \mathbf{P}(\alpha \wedge (\beta \wedge \gamma))$
(CD)	$\mathbf{P}(\alpha \wedge (\beta \vee \gamma)) \leftrightarrow \mathbf{P}((\alpha \wedge \beta) \vee (\alpha \wedge \gamma))$
Rules	
(MP)	Given φ and $(\varphi \rightarrow \psi)$, infer ψ.

System $\mathbb{DLSP}^s$

	System $\mathbb{DLSP}$ with the following axiom
(EXP)	$\mathbf{P}(p_1 \wedge ... \wedge p_k) \rightarrow \mathbf{P}(m_1 \cdot p_1 \wedge ... \wedge m_k \cdot p_k)$

As a demonstration of our systems, we show that

Proposition 2.8 DC1 : $\mathbf{P}((\alpha \vee \beta) \wedge (\alpha \vee \gamma)) \rightarrow \mathbf{P}(\alpha \vee (\beta \wedge \gamma))$ *is provable in* $\mathbb{DLSP}$ *and in* $\mathbb{DLSP}^s$*.*

Proof. The derivation goes as follows:

1.	$\mathbf{P}((\alpha \vee \beta) \wedge (\alpha \vee \gamma)) \rightarrow \mathbf{P}(((\alpha \vee \beta) \wedge \alpha) \vee ((\alpha \vee \beta) \wedge \gamma))$	(CD)
2.	$\mathbf{P}(((\alpha \vee \beta) \wedge \alpha) \vee ((\alpha \vee \beta) \wedge \gamma)) \rightarrow \mathbf{P}((\alpha \vee \beta) \wedge \gamma)$	(FC)
3.	$\mathbf{P}((\alpha \vee \beta) \wedge \gamma) \rightarrow \mathbf{P}(\gamma \wedge (\alpha \vee \beta))$	($\mathtt{COM}_\wedge$)
4.	$\mathbf{P}(\gamma \wedge (\alpha \vee \beta)) \rightarrow \mathbf{P}((\gamma \wedge \alpha) \vee (\gamma \wedge \beta))$	(CD)
5.	$\mathbf{P}((\gamma \wedge \alpha) \vee (\gamma \wedge \beta)) \rightarrow (\mathbf{P}(\gamma \wedge \alpha) \wedge \mathbf{P}(\gamma \wedge \beta))$	(FC)
6.	$\mathbf{P}((\alpha \vee \beta) \wedge (\alpha \vee \gamma)) \rightarrow (\mathbf{P}(\gamma \wedge \alpha) \wedge \mathbf{P}(\gamma \wedge \beta))$	(1−5, TAUT, MP)
7.	$\mathbf{P}(\gamma \wedge \alpha) \rightarrow \mathbf{P}\alpha$	(CE)
8.	$\mathbf{P}(\gamma \wedge \beta) \rightarrow \mathbf{P}(\beta \wedge \gamma)$	($\mathtt{COM}_\wedge$)
9.	$(\mathbf{P}(\gamma \wedge \alpha) \wedge \mathbf{P}(\gamma \wedge \beta)) \rightarrow (\mathbf{P}\alpha \wedge \mathbf{P}(\beta \wedge \gamma))$	(7, 8, TAUT, MP)
10.	$(\mathbf{P}\alpha \wedge \mathbf{P}(\beta \wedge \gamma)) \rightarrow \mathbf{P}(\alpha \vee (\beta \wedge \gamma))$	(FC)
11	$(\mathbf{P}(\gamma \wedge \alpha) \wedge \mathbf{P}(\gamma \wedge \beta)) \rightarrow \mathbf{P}(\alpha \vee (\beta \wedge \gamma))$	(9, 10, TAUT, MP)
12.	$\mathbf{P}((\alpha \vee \beta) \wedge (\alpha \vee \gamma)) \rightarrow \mathbf{P}(\alpha \vee (\beta \wedge \gamma))$	(6, 11, TAUT, MP)

□

Based on the Propositions 2.1, 2.5, 2.6, 2.7, and the fact that the connectives outside **P** are classical in the semantics, we can show:

Theorem 2.9 (Soundness Theorem) $\mathbb{DLSP}$ *is sound w.r.t. the class of all deontic models, and* $\mathbb{DLSP}^s$ *is sound w.r.t. the class of singleton deontic models.*

The completeness proof will make use of a normal form of **DLSP**, towards which we first present an example below.

Example 2.10 We use axioms in $\mathbb{DLSP}$ to equivalently transform the following formula into a conjunction of formulas in the shape of $\mathbf{P}(p_1 \wedge ... \wedge p_n)$.

$$\mathbf{P}(p_1 \vee (p_2 \wedge ((p_3 \vee p_4) \wedge p_5))).$$

The formula is logically equivalent to

1.	$\mathbf{P}p_1 \wedge \mathbf{P}(p_2 \wedge ((p_3 \vee p_4) \wedge p_5))$	(FC)
2.	$\mathbf{P}p_1 \wedge \mathbf{P}((p_5 \wedge p_2) \wedge (p_3 \vee p_4))$	$(\texttt{ASSO}_\wedge + \texttt{COM}_\wedge)$
3.	$\mathbf{P}p_1 \wedge \mathbf{P}(((p_5 \wedge p_2) \wedge p_3) \vee ((p_5 \wedge p_2) \wedge p_4))$	(CD)
4.	$\mathbf{P}p_1 \wedge \mathbf{P}((p_5 \wedge p_2) \wedge p_3) \wedge \mathbf{P}((p_5 \wedge p_2) \wedge p_4)$	(FC)

In fact, such a method can be uniformly used to transform any formula of the form $\mathbf{P}\alpha$ into a similar form.

Lemma 2.11 (Normal Form for $\mathbf{P}\alpha$) *For any $\alpha \in$ **AT**, $\mathbf{P}\alpha$ is logically equivalent to a formula of the form $(\mathbf{P}\beta_1 \wedge ... \wedge \mathbf{P}\beta_k)$ where for each $1 \leq i \leq k$, β_i is in the shape of $\mathbf{P}(p_1 \wedge ... \wedge p_n)$, which is called a **normal form** for $\mathbf{P}\alpha$.*[6]

Proof. (Sketch) We follow the same procedure in the example above to transform any $\mathbf{P}\alpha$ into a normal form. If α is in the shape of $p_1 \wedge ... \wedge p_n$, then it trivially holds. If not, then use (FC) and (CD) alternately with the help of $(\texttt{COM}_\wedge)$ and $(\texttt{ASSO}_\wedge)$ to split action types and eliminate disjunction symbols. □

Now we can rewrite **DLSP** into a new form.

Proposition 2.12 *For any formula $\varphi \in$ **DLSP**, φ is logically equivalent to a formula in the following language (denoted by **DLSP***):*

$$::= \bot \mid p \mid \mathbf{P}(p_1 \wedge ... \wedge p_n) \mid \neg\psi \mid (\psi \wedge \psi) \mid (\psi \vee \psi) \mid (\psi \rightarrow \psi),$$

where $p, p_1, ..., p_n \in P$.

Qua expressive power, **DLSP*** and **DLSP** are the same language, but in **DLSP***, no disjunction symbol occurs within **P**, which will facilitate our canonical model construction in the next section.

3 Completeness

In this section, we prove strong completeness results for $\mathbb{DLSP}$ and $\mathbb{DLSP}^*$.

3.1 System $\mathbb{DLSP}$

Note that due to the validity of $\texttt{ASSO}_\wedge$ and $\texttt{COM}_\wedge$, we will treat an action token of type $(p_1 \wedge ... \wedge p_n)$ as an n-ary tuple of action tokens modulo paring.

Definition 7 (All-Distinct Token) *An action token of type $(p_1 \wedge ... \wedge p_n)$ is* all-distinct *if tokens of the same atomic action type in the tuple are pairwise distinct.*

Now let Σ be a maximally $\mathbb{DLSP}$-consistent set of **DLSP*** formulas (MCS). To build a canonical model $\mathcal{M}^C_\Sigma$, we first construct a canonical action space S^C_Σ. The basic idea is that for each $p \in P$, we set S^C_Σ to be the set of n distinct p^i symbols where $n \in \mathbb{N}_{>0}$ is the least number such that $\neg\mathbf{P}(n \cdot p) \in \Sigma$ (cf. Proposition 2.7 for the "$n \cdot p$" notation).

[6] Due to propositional tautologies for $\wedge$, permutations do not matter in the normal form.

Definition 8 (Canonical Action Space) *Given Σ, we define S^C_Σ by distinguishing the two cases of $p \in P$:*

- *If there is an $i \in \mathbb{N}_{>0}$ such that the formula $\neg\mathbf{P}(i \cdot p) \in \Sigma$, assume that n is the least of such i, and let $S^C_\Sigma(p) := \{p^1, p^2, ..., p^n\}$, in which each p^j is the propositional letter p superscript with the numeral j.*
- *If not, i.e., $\mathbf{P}(i \cdot p) \in \Sigma$ for all $i \in \mathbb{N}_{>0}$, let $S^C_\Sigma(p) := \{p^1, p^2, ...\}$.*

For composite $\alpha \in \mathbf{AT}$, we define $S^C_\Sigma(\alpha)$ recursively as in the definition of S.

Note that for distinct $p, q \in P$, $S^C_\Sigma(p) \cap S^C_\Sigma(q) = \emptyset$. The following lemma plays an important role in the later proofs.

Lemma 3.1 *For any formula φ of the form $\mathbf{P}(m_1 \cdot p_{t_1} \wedge ... \wedge m_k \cdot p_{t_k})$ where p_{t_i}, p_{t_j} are pairwise distinct, if $\varphi \in \Sigma$, then for any $1 \leq j \leq k$, $m_j < |S^C_\Sigma(p_{t_j})|$.*

Proof. Assume that $\varphi \in \Sigma$ and suppose (towards a contradiction) that there is $1 \leq j \leq k$ such that $m_j \geq |S^C_\Sigma(p_{t_j})|$. Thus $|S^C_\Sigma(p_{t_j})|$ is finite, say $|S^C_\Sigma(p_{t_j})| = n$. By definition, $\neg\mathbf{P}(n \cdot p_{t_j}) \in \Sigma$. However, since $\varphi \in \Sigma$ and $m_j \geq n$, by (CE) and (MP), $\mathbf{P}(n \cdot p_{t_j}) \in \Sigma$. That is a contradiction. □

This lemma shows the size of the action space is *more than enough* to guarantee the existence of all-distinct action tokens of the type $(m_1 \cdot p_{t_1} \wedge ... \wedge m_k \cdot p_{t_k})$ when $\mathbf{P}(m_1 \cdot p_{t_1} \wedge ... \wedge m_k \cdot p_{t_k}) \in \Sigma$. Based on this and S^C_Σ we will build a pointed deontic model $\mathcal{M}^C_\Sigma, w$ such that the truth lemma holds, i.e.,$\mathbf{P}(p_1 \wedge ... \wedge p_n) \in \Sigma$ iff $\mathcal{M}^C_\Sigma, w \vDash \mathbf{P}(p_1 \wedge ... \wedge p_n)$. The idea is simple: given a designated world w, build the accessible worlds according to formulas $\mathbf{P}(m_1 \cdot p_{t_1} \wedge ... \wedge m_k \cdot p_{t_k}) \in \Sigma$. The subtlety is that we should only realize action tokens that are *necessary* to witness the truth of those φ, but no more, for we also need tokens not realizable to witness $\neg\mathbf{P}(p_1 \wedge ... \wedge p_n) \in \Sigma$. The later task is doable because we have some spare tokens in S^C_Σ based on Lemma 3.1.

To define $\mathcal{M}^C_\Sigma$, we first fix an ordering of (countably many) propositional letters $p_0, p_1, p_2, ...$ For formulas $\mathbf{P}(p_{s_1} \wedge ... \wedge p_{s_n}) \in \Sigma$, note that based on $\texttt{ASSO}_\wedge$ and $\texttt{COM}_\wedge$, we only need to consider $\mathbf{P}(m_1 \cdot p_{t_1} \wedge ... \wedge m_k \cdot p_{t_k}) \in \Sigma$ such that p_{t_i} and p_{t_j} are distinct and ordered according to the order of propositional letters, e.g., $\mathbf{P}(3 \cdot p_2 \wedge 4 \cdot p_6)$. We propose a simple representation for such formulas using functions on natural numbers.

Definition 9 *For any φ of the form $\mathbf{P}(m_1 \cdot p_{t_1} \wedge ... \wedge m_k \cdot p_{t_k}) \in \Sigma$ such that p_{t_i} and p_{t_j} are distinct and ordered according to the order of propositional letters, we define $f_\varphi : \mathbb{N} \to \mathbb{N}$ such that*

$$f_\varphi(i) = \begin{cases} m_j & i = t_j \text{ for some } 1 \leq j \leq k; \\ 0 & i \neq t_j \text{ for any } 1 \leq j \leq k. \end{cases}$$

For example, $\mathbf{P}(3 \cdot p_2 \wedge 4 \cdot p_6)$ is represented by the function f such that $f(2) = 3$, $f(6) = 4$ and $f(i) = 0$ for any $i \in \mathbb{N} \setminus \{2, 6\}$. We collect these (countably many) functions in F_Σ.

Definition 10 *For any $f \in F_\Sigma$, we define $G_f :=$*

$$\{g : \mathbb{N} \to \mathcal{P}(\bigcup_{p \in P} (S^C_\Sigma(p))) \mid \textit{ for any } i \in \mathbb{N}, g(i) \subseteq S^C_\Sigma(p_i) \textit{ and } |g(i)| = f(i)\}.$$

Intuitively, each $g \in G_f$ assigns a subset of the canonical action space of each p_i whose cardinality is $f(i)$. It follows if $f(i) = 0$ then $g(i) = \emptyset$. Let $G_\Sigma = \bigcup\{G_f \mid f \in F_\Sigma\}$, which will be used to build the canonical model below. We first have the following observation:

Proposition 3.2 *Given a MCS Σ and any distinct $f, f' \in F_\Sigma$, we have: (1) G_f is not empty; (2) $G_f \cap G_{f'} = \emptyset$.*

Proof. For (1), by Lemma 3.1, given $f \in F_\Sigma$, $f(i) < |S^C_\Sigma(p_i)|$, thus each G_f is not empty. For (2), note that if $f \neq f'$ then there must be some i such that $f(i) \neq f'(i)$ and this will distinguish any $g \in G_f$ from any $g' \in G_{f'}$ due to the cardinality requirement $|g(i)| = f(i) \neq f'(i) = |g'(i)|$. □

Definition 11 (Canonical Deontic Model) *Given a MCS Σ, we define the model $\mathcal{M}^C_\Sigma = (S^C_\Sigma, W^C, R^C, A^C)$ where:*

- $W^C = \{w\} \cup G_\Sigma$; $R^C = \{(w, g) \mid g \in G_\Sigma\}$;
- $A^C(p_i, u) = \begin{cases} S^C_\Sigma(p_i) & \textit{if } u = w \textit{ and } p_i \in \Sigma, \\ \emptyset & \textit{if } u = w \textit{ and } p_i \notin \Sigma, \\ u(i) & \textit{if } u \in G_\Sigma; \end{cases}$

and $A^C(\alpha, u)$ for composite α is defined as in Definition 5.

By Proposition 3.2 (2), if $g \in G_\Sigma$ then there is a unique $f \in F_\Sigma$ s.t. $g \in G_f$. Intuitively, each $g \in G_f$ realizes some all-distinct token (cf. Definition 7) of the formula $\mathbf{P}\alpha \in \Sigma$ corresponding to f, and G_f realize all the necessary tokens.

Lemma 3.3 (Truth Lemma for $\mathbb{DLSP}$) *Let Σ be a maximally $\mathbb{DLSP}$-consistent set of* **DLSP*** *formulas. For any $\varphi \in \Sigma$,*

$$\mathcal{M}^C_\Sigma, w \vDash \varphi \iff \varphi \in \Sigma.$$

Proof. We show this by induction on the structure of φ.

- $\varphi = \bot$: by definition, $\mathcal{M}^C_\Sigma, w \nvDash \bot$. And, since Σ is consistent, $\bot \notin \Sigma$. So, it trivially holds.
- $\varphi = p$: $\mathcal{M}^C_\Sigma, w \vDash p$ iff $A^C(p, w) \neq \emptyset$ iff $A^C(p, w) = S^C_\Sigma(p)$ iff $p \in \Sigma$.
- $\varphi = \neg\psi$, $(\psi \wedge \chi)$, $(\psi \vee \chi)$ or $(\psi \to \chi)$: it holds by definition, inductive hypothesis and maximality of Σ.
- $\varphi = \mathbf{P}(p_1 \wedge ... \wedge p_n)$: By $(\mathtt{COM}_\wedge)$ and $(\mathtt{ASSO}_\wedge)$, φ is logically equivalent to a formula ψ of the form $\mathbf{P}(m_1 \cdot p_{t_1} \wedge ... \wedge m_k \cdot p_{t_k})$ where p_{t_i} and p_{t_j} are pairwise distinct and ordered. We only need to show that

$$\mathcal{M}^C_\Sigma, w \vDash \psi \iff \psi \in \Sigma.$$

$\Leftarrow$: Assume that $\psi \in \Sigma$. We have the corresponding $f_\psi \in F_\Sigma$. We show that each token in $S^C_\Sigma(m_1 \cdot p_{t_1} \wedge ... \wedge m_k \cdot p_{t_k})$ is realized in some $g \in G_{f_\psi}$. Take an arbitrary token x of $(m_1 \cdot p_{t_1} \wedge ... \wedge m_k \cdot p_{t_k})$, let function h be defined by setting $h(i) = \{p^j_i \mid p^j_i \text{ appear in } x\} \subseteq S^C_\Sigma(p_i)$. By Proposition 3.2, it is clear that there is a $g \in G_{f_\psi}$ such that $h(i) \subseteq g(i)$ for all $i \in \mathbb{N}$. Therefore by the definition of A^C, $x \in A^C((m_1 \cdot p_{t_1} \wedge ... \wedge m_k \cdot p_{t_k}), g)$.
$\Rightarrow$: Assume that $\quad \notin \Sigma$. To show $\mathcal{M}^C_\Sigma, w \not\models \psi$, we need to find some token in $S^C_\Sigma(m_1 \cdot p_{t_1} \wedge ... \wedge m_k \cdot p_{t_k})$ cannot be witnessed by any successor. The crucial point here is that our definition of S^C_Σ and A^C together guarantee that some action tokens are indeed **left out** at every $g \in G_\Sigma$.

Now we consider two cases:

- ○ If for any $1 \leq j \leq k$, $m_j \leq |S^C_\Sigma(p_{t_j})|$, we take an all-distinct token $x \in S(m_1 \cdot p_{t_1} \wedge ... \wedge m_k \cdot p_{t_k})$ and show it is not realizable in G_Σ, thus $\mathcal{M}^C_\Sigma, w \nvDash \psi$. Suppose not, then there is a $g \in G_\Sigma$ that realizes x, then there is a unique f such that $g \in G_f$ by Proposition 3.2. Since g realizes all-distinct token x, then we have $f(t_j) = |g(t_j)| = |A^C(p_{t_j}, g)| \geq m_j$ for any $1 \leq j \leq k$. Due to our construction, there must be a $\chi \in \Sigma$ such that $f = f_\chi$. Therefore, χ must be of the form $\mathbf{P}((m'_1 \cdot p_{t_1} \wedge ... \wedge m'_k \cdot p_{t_k}) \wedge (m'_{k+1} \cdot p_{t_{k+1}} \wedge ... \wedge m'_{k+l} \cdot p_{t_{k+l}})) \in \Sigma$ such that $m'_j = f(t_j) \geq m_j$. By (`CE`) and (`MP`), $\psi \in \Sigma$, contradicting to the assumption that $\psi \notin \Sigma$. So, $\mathcal{M}^C_\Sigma, w \nvDash \psi$.
- ○ If there is $1 \leq j \leq k$ such that $m_j > |S^C_\Sigma(p_{t_j})|$, thus $S^C_\Sigma(p_{t_j})$ is finite, say $|S^C_\Sigma(p_{t_j})| = n$. Suppose towards a contradiction that $\mathcal{M}^C_\Sigma, w \vDash \psi$. Thus by the validity of `CE`, $\mathcal{M}^C_\Sigma, w \vDash \mathbf{P}(n \cdot p_{t_j})$. Hence, to realize the token using all the atomic tokens in $S^C_\Sigma(p_{t_j})$, there must be a $g \in G$ such that $A^C(p_{t_j}, g) = g(t_j) = S^C_\Sigma(p_{t_j})$. Note that by Proposition 3.2 there must be a unique f such that $g \in G_f$ and $f(t_j) = |g(t_j)| = |S^C_\Sigma(p_{t_j})| = n$. Therefore there is a $\chi \in \Sigma$ such that $f = f_\chi$. However this means χ must be in the shape of $\mathbf{P}(n \cdot p_{t_j} \wedge \beta) \in \Sigma$. By (`CE`), $\mathbf{P}(n \cdot p_{t_j}) \in \Sigma$ contradicting to the fact that $|S^C_\Sigma(p_{t_j})| = n$ (cf. Definition 8). Therefore $\mathcal{M}^C_\Sigma, w \nvDash \psi$. □

Based on the truth lemma, by a Lindenbaum-like argument, we can show:

Theorem 3.4 (Completeness Theorem for $\mathbb{DLSP}$) $\mathbb{DLSP}$ *is strongly complete with respect to the class of all deontic models.*

Proof. We show its contraposition that every $\mathbb{DLSP}$-consistent set of formulas is satisfiable by the canonical model. First, using a Lindenbaum-like argument to extend a $\mathbb{DLSP}$-consistent set into an MCS Σ, and then turn this set into a provably equivalent set Σ' of $\mathbb{DLSP}^*$ formulas by using the normal form. Finally, we apply the truth lemma to construct the model for Σ' (thus also satisfies Σ). □

Note that $\mathbb{DLSP}$ is also complete over all *serial* models, i.e., the models where every node has a successor. The canonical model in Definition 11 works as before, except for (1) we need to add self-loops to $g \in G_\Sigma$ and (2) in case of $\{\neg\mathbf{P}p \mid p \in P\} \subseteq \Sigma$ thus $G_\Sigma = \emptyset$, we need to add a self-loop successor to w

where no tokens are realized.

3.2 System $\mathbb{DLSP}^s$

Now let Σ be a maximally $\mathbb{DLSP}^s$-consistent set of $\mathbf{DLSP}^*$ formulas. First we define the singleton action space S^s_Σ straightforwardly.

Definition 12 (Canonical Singleton Action Space) *Given Σ, we define the canonical singleton action space S^s_Σ such that $S^s_\Sigma(p) := \{p\}$ for any $p \in P$ and $S^s_\Sigma(\alpha)$ is defined recursively as above for any composite $\alpha \in \mathbf{AT}$.*

To define the canonical singleton deontic model $\mathcal{M}^s_\Sigma = (S^s_\Sigma, W^s, R^s, A^s)$, we essentially apply the same method as we use to define the canonical deontic model in Definition 11. In fact, the definition is simpler here since $\mathbf{P}(m_1 \cdot p_{t_1} \wedge ... \wedge m_k \cdot p_{t_k})$ is logically equivalent to $\mathbf{P}(p_{t_1} \wedge ... \wedge p_{t_k})$ by extra validities (EXP) in $\mathbb{DLSP}^s$ where p_{t_i} and p_{t_j} are distinct and ordered. Thus we only consider formulas φ of the latter form in Σ and define f_φ, G_{f_φ} as before. So, note here that for any $i \in \mathbb{N}$, if $i = t_j$, then $f_\varphi(i) = 1$, and otherwise $f_\varphi(i) = 0$. And by this feature of f_φ and S^s_Σ as singleton action space, there is indeed a unique $g \in G_{f_\varphi}$ that $g(i) = \{p_{t_j}\}$ if $i = t_j$ and $g(i) = \emptyset$ otherwise. We collect all such g in G'_Σ.

Definition 13 (Canonical Singleton Deontic Model) *Given Σ, we define the singleton deontic model $\mathcal{M}^s_\Sigma = (S^s_\Sigma, W^s, R^s, A^s)$ where:*

- $W^s = \{v\} \cup G'_\Sigma$; $R^s = \{(v, g) \mid g \in G'_\Sigma\}$;
- $A^s(p_i, u) = \begin{cases} S^s_\Sigma(p_i) & \textit{if } u = v \textit{ and } p_i \in \Sigma, \\ \emptyset & \textit{if } u = v \textit{ and } p_i \notin \Sigma, \\ u(i) & \textit{if } u \in G'_\Sigma; \end{cases}$

and $A^s(\alpha, u)$ for composite α is defined as in Definition 5.

Lemma 3.5 (Truth Lemma for $\mathbb{DLSP}^s$) *Let Σ be a maximally $\mathbb{DLSP}^s$-consistent set of* $\mathbf{DLSP}^*$ *formulas. For any $\varphi \in \Sigma$,*

$$\mathcal{M}^s_\Sigma, v \vDash_s \varphi \Longleftrightarrow \varphi \in \Sigma.$$

Proof. Similarly by induction on the structure of φ, we only show the case where $\varphi = \mathbf{P}(p_1 \wedge ... \wedge p_n)$. By the validities of (EXP) and (CE), φ is logically equivalent to $\psi = \mathbf{P}(p_{t_1} \wedge ... \wedge p_{t_k})$ where $p_{t_1}, ..., p_{t_k} \in P$ are pairwise distinct and ordered. We show this inductive case holds for ψ.
$\Leftarrow$: Assume that $\mathbf{P}(p_{t_1} \wedge ... \wedge p_{t_k}) \in \Sigma$. Then the corresponding f_ψ is defined. By the definition of $\mathcal{M}^s_\Sigma$, there is a unique $g \in G_{f_\psi}$ such that for any $1 \leq j \leq k$, $A^s(p_{t_j}, g) = g(t_j) = \{p_{t_j}\}$. Hence, $A^s(p_{t_1} \wedge ... \wedge p_{t_k}, g) = \{(p_{t_1}, ..., p_{t_k})\} = S^s_\Sigma(p_{t_1} \wedge ... \wedge p_{t_k})$. Therefore, $\mathcal{M}^s_\Sigma, v \vDash \mathbf{P}(p_{t_1} \wedge ... \wedge p_{t_k})$.
$\Rightarrow$: Assume that $\mathcal{M}^s_\Sigma, v \vDash \mathbf{P}(p_{t_1} \wedge ... \wedge p_{t_k})$. Since $S^s_\Sigma(p_{t_1} \wedge ... \wedge p_{t_k}) = \{(p_{t_1}, ..., p_{t_k})\}$, there is $g \in G'_\Sigma$ such that for any $1 \leq j \leq k$, $A^s(p_{t_j}, g) = \{p_{t_j}\}$. So, by the definition of G'_Σ, there is φ of the form $\mathbf{P}(p_{t_1} \wedge ... \wedge p_{t_k} \wedge ... \wedge p_{t_{k+l}})$ where p_i and p_j are pairwise distinct and ordered such that $G_{f_\varphi} = \{g\}$. Hence, by (CE) and (MP), $\mathbf{P}(p_1 \wedge ... \wedge p_n) \in \Sigma$. □

Theorem 3.6 (Completeness Theorem for $\mathbb{DLSP}^s$) *$\mathbb{DLSP}^s$ is strongly complete with respect to the class of all singleton deontic models.*

4 Generalizations

In our framework, the function A in the model computes the executed tokens of complex types based on atomic tokens in each world. Such a setting may raise the question of whether there are too many tokens "automatically" executed on each world. Our framework can be generalized if we can take a subset of those generated tokens as the executed ones. In this section, we explore this possibility in a specific setting where the tuples of executed tokens might not be executed, e.g., the realization of token a of type α and token b of type β on a world does not necessarily imply the realization of (a, b) of type $\alpha \wedge \beta$, and the realization of other tokens such as (a, b, a, a, b, a), due to limited executive resources. Surprisingly, this change does not affect the logic much.

Given an action space S, we define $I_S := \bigcup_{\alpha \in \mathbf{AT}} S(\alpha)$. That is, I_S collects exactly all possible action tokens of all action types under S.

Definition 14 (Closure Set of Tokens) *We say a set $T \subseteq I_S$ of action tokens is* closed *iff*

(i) *$a \in T$ if and only if $(a, 0) \in T$ if and only if $(a, 1) \in T$;*

(ii) *If $a, b \in I_S$ and $(a, b) \in T$, then $a \in T$ and $b \in T$;*

(iii) *If $a, b \in I_S$ and $(a, b) \in T$, then $(b, a) \in T$;*

(iv) *If $a, b, c \in I_S$, then $((a, b), c) \in T$ if and only if $(a, (b, c)) \in T$;*

(v) *If $a, b \in I_S$, then $((a, b), i) \in T$ if and only if $(a, (b, i)) \in T$, $i \in \{0, 1\}$.*

These closure properties correspond to the basic axioms in our system $\mathbb{DLSP}$. However, it is possible for $a, b \in T$ but $(a, b) \notin T$. Below we make use of closure sets to introduce two ways of generalizing our model and semantics. For the first one, an extra function σ that assigns to each world a closure set of tokens is straightforwardly added into a deontic model.

Definition 15 (I-Type General Deontic Model) *A I-type general deontic model $\mathcal{M}^G$ is a 5-ary tuple (S, W, R, A, σ) such that*

- *(S, W, R, A) is a deontic model.*
- *$\sigma : W \to \wp(I_S)$ such that for any $w \in W$, $\sigma(w)$ is closed.*

Given a I-type model, we further ascertain for each action type α, what action tokens are done at each possible world w by the following binary function A^G:

$$A^G(\alpha, w) := A(\alpha, w) \cap \sigma(w).$$

At each world, only those action tokens in the closure set given by σ are *executable.* So the original binary function A just computes what tokens are *in principle* executed; rather, what are *actually* executed are those given by the defined A^G. For another generalizing way, we omit the original A and add such σ in a deontic model.

Definition 16 (II-Type General Deontic Model) *A II-type general deontic model is a 4-ary tuple (S, W, R, σ) such that*

- *S, W, R are as usual,*
- *$\sigma : W \to \wp(I_S)$ such that for any $w \in W$, $\sigma(w)$ is closed.*

As the case in I-type general deontic model, we also ascertain a binary function A^G rather by $A^G(\alpha, w) := S(\alpha) \cap \sigma(w)$.

Here, what action tokens are executed are directly given by σ, and A^G further shows what tokens are executed for each action type. Now we define semantics for any i-type general deontic model $\mathcal{M}^G$ by replacing all appearances of binary function A with A^G in clauses of Definition 6, i.e. $\mathcal{M}, w \Vdash_i p \iff A^G(p, w) \neq \emptyset$ and for any $a \in S(\alpha)$, there is a $v \in W$ such that wRv and $a \in A^G(\alpha, v)$, where $i \in \{\mathrm{I}, \mathrm{II}\}$.

Obviously, a II-type general deontic model is a special kind of I-type model where $A(\alpha, w) = S(\alpha)$ for any $\alpha \in \mathbf{AT}$ and $w \in W$. In fact, a I-type model can also be transformed into a II-type one under certain conditions while truths are preserved. We present the relevant definitions and results below.

Definition 17 (Disjointed Action Space) *Given a non-empty set I of action tokens such that $I \cap \{0, 1\} = \emptyset$, a disjointed action space S is an action space such that for any $p, q \in P$, if $p \neq q$, then $S(p) \cap S(q) = \emptyset$.*

Definition 18 (Co-Executive) *For any binary function A in a (I-type general) deontic model $\mathcal{M} = (S, W, R, A)$ (or (S, W, R, A, σ)), we say that A is co-executive ("Co") if for any $x \in I_S$ and $w \in W$, if $x \in S(\alpha) \cap S(\beta)$, then $x \in A(\alpha, w) \Leftrightarrow x \in A(\beta, w)$.*

Lemma 4.1 *For any (I-type general) deontic model based on a disjointed action space, A is co-executive.*

Proposition 4.2 *Given a disjointed action space S, for each I-type general deontic model $\mathcal{M} = (S, W, R, A, \sigma)$, there is a II-type general deontic model $\mathcal{M}' = (S, W, R, \sigma')$ such that for any $\varphi \in$ **DLSP** and $w \in W$, $\mathcal{M}, w \Vdash_{\mathrm{I}} \varphi$ if and only if $\mathcal{M}', w \Vdash_{\mathrm{II}} \varphi$; and vice versa.*

Therefore, based on disjointed action spaces, I-type and II-type general deontic model can be transformed into each other while preserving truths and thus treated together as a "general deontic model". Now under the class of general deontic models, first we can show that:

Theorem 4.3 *$\mathbb{DLSP}$ is sound with respect to the class of all general models.*

We can adapt the canonical model (based on a disjoint action space) in Definition 11 to show the following completeness, which shows our logic cannot differentiate whether the closure of conjunctive tokens is imposed.

Theorem 4.4 *$\mathbb{DLSP}$ is strongly complete over the class of all general models.*

Following similar ideas, we can summarize the results of complete axiom systems under different kinds of deontic models:

Deontic Models	A	A: Co	(A, σ)	(A, σ): Co	σ
S	$\mathbb{DLSP}$	$\mathbb{DLSP}$	$\mathbb{DLSP}$	$\mathbb{DLSP}$	$\mathbb{DLSP}$
S: Disjointed	$\mathbb{DLSP}$	$\mathbb{DLSP}$	$\mathbb{DLSP}$	$\mathbb{DLSP}$	$\mathbb{DLSP}$
S: Singleton	$\mathbb{DLSP}^s$	$\mathbb{DLSP}^s$	$\mathbb{DLSP}$	$\mathbb{DLSP}$	$\mathbb{DLSP}$
S: Singleton, Disjointed	$\mathbb{DLSP}^s$	$\mathbb{DLSP}^s$	$\mathbb{DLSP}$	$\mathbb{DLSP}$	$\mathbb{DLSP}$

In the future, we may relax other closure properties in Definition 14.

5 Conclusions and Future Work

In this paper, we have taken the first steps towards treating strong permission as a $\forall\Diamond$ bundle in deontic logic, drawing on a BHK-like semantics to assign tokens to action types. Through our semantics, we uncover a new invalid law that aligns with our linguistic intuition. The failure of this particular distribution law, `DCr`, sets our framework apart from most other approaches that admit free choice permission. Another invalid law within our framework is the duplication law `EX` within the permission operator. We demonstrated that, when combined with the widely acceptable distribution law `CD` and the free choice law `FC`, this seemingly innocuous law leads to the counter-intuitive consequence that $\mathbf{P}(\alpha \vee \beta) \to \mathbf{P}(\alpha \wedge \beta)$. The invalidity of `EX` in our setting arises from the fact that permitting various tokens of one type does not imply permitting their arbitrary combinations. We also present cases where the action spaces are singletons, resulting in a restricted version `EXP` of `EX`. To ensure that we do not overlook any other (in)validities, we completely axiomatize the logics over arbitrary models and singleton models, obtaining intuitive proof systems $\mathbb{DLSP}$ and $\mathbb{DLSP}^s$. We also generalize our framework to make it possible that the tokens are not closed under conjunctions due to resource bounds. It turns out the logic will stay almost the same in this more general setting.

There are many further directions. First of all, the generalizations of the semantics in Section 4 give us the fine-grain control needed on the semantics w.r.t. validity of the axioms. Thus it is possible to address further subtleties regarding the law of permissions, such as making `CE` invalid in certain contexts (only the combination of two actions is permitted without allowing single ones). It remains to see whether we can handle the concerns in [2] about the purely semantic approaches. We will also extend the framework with other connectives, such as $\neg$ and $\to$, and modalities, such as $\mathbf{O}$ and $\mathbf{F}$, enriching the current formal treatment. Comparisons and connections with other approaches where action types and tokens are interpreted by similar inquisitive semantics like [1,3] are also worth exploring.

Acknowledgment The authors thank Frank Veltman for his detailed comments on an earlier version of the paper, which helped to improve the presentation of the paper. The critical comments from the anonymous reviewers of DEON2023 are also helpful in making the paper more accessible. The authors also would like to thank Maria Aloni, Jialiang Yan, Fengkui Ju, and Chenwei Shi for their comments and relevant pointers. The work is supported by NSSF grant 19BZX135.

References

[1] Aloni, M., *Free choice, modals, and imperatives*, Natural Language Semantics **15** (2007), p. 65–94.

[2] Aloni, M., *Logic and conversation: the case of free choice*, Semantics and Pragmatics **15** (2022), pp. 5–EA.

[3] Aloni, M. and I. Ciardelli, *A logical account of free-choice imperatives*, The dynamic, inquisitive, and visionary life of φ (2013), pp. 1–17.

[4] Anglberger, A. J., N. Gratzl and O. Roy, *Obligation, free choice, and the logic of weakest permissions*, The Review of Symbolic Logic **8** (2015), pp. 807–827.

[5] Anglberger, A. J. J., H. Dong and O. Roy, *Open reading without free choice*, in: F. Cariani, D. Grossi, J. Meheus and X. Parent, editors, *Deontic Logic and Normative Systems*, Lecture Notes in Computer Science (2014), p. 19–32.

[6] Bentzen, M. M., *Action type deontic logic*, Journal of Logic, Language and Information **23** (2014), pp. 397–414.

[7] Broersen, J., *Action negation and alternative reductions for dynamic deontic logics*, Journal of applied logic **2** (2004), pp. 153–168.

[8] Castro, P. F. and P. Kulicki, *Deontic logics based on boolean algebra*, in: R. Trypuz, editor, *Krister Segerberg on Logic of Actions*, Springer Netherlands, Dordrecht, 2014 p. 85–117.

[9] Chellas, B. F., “Modal Logic: An Introduction,” Cambridge University Press, 1980.

[10] Dong, H. and O. Roy, *Three deontic logics for rational agency in games*, Studies in Logic **8** (2015), pp. 7–31.

[11] Governatori, G. and A. Rotolo, *Is free choice permission admissible in classical deontic logic?*, in: *Proceedings Deontic Logic in Computer Science (DEON 2020)*, 2020, pp. 255–271.

[12] Hansson, S. O., *The varieties of permission*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems, vol. 1*, College publications, London, 2013 pp. 195–240.

[13] Hintikka, J., *Some main problems of deontic logic*, in: R. Hilpinen, editor, *Deontic Logic: Introductory and Systematic Readings*, Springer Netherlands, Dordrecht, 1971 pp. 59–104.

[14] Kamp, H., *Free choice permission*, Proceedings of the Aristotelian Society **74** (1973), pp. 57–74.

[15] Kolmogorov, A., *Zur deutung der intuitionistischen logik*, Mathematische Zeitschrift (1932), pp. 58–65.

[16] McNamara, P., *Deontic logic*, in: *Handbook of the History of Logic*, Elsevier, 2006 pp. 197–288.

[17] Medvedev, Y. T., *Interpretation of logical formulas by means of finite problems*, Dokl. Akad. Nauk SSSR **169** (1966), pp. 20–23.

[18] Meyer, J. J. C., *A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic*, Notre Dame Journal of Formal Logic **29** (1987), pp. 109–136.

[19] Segerberg, K., *A deontic logic of action*, Studia logica **41** (1982), pp. 269–282.

[20] Trypuz, R. and P. Kulicki, *On deontic action logics based on boolean algebra*, Journal of Logic and Computation **25** (2015), pp. 1241–1260.

[21] van Benthem, J., *Minimal deontic logics*, Bulletin of the Section of Logic **8** (1979), pp. 36–42.

[22] Wang, H., Y. Wang and Y. Wang, *Inquisitive logic as an epistemic logic of knowing how*, Annals of Pure and Applied Logic (2022), p. 103145.

[23] Wang, Y., *A logic of goal-directed knowing how*, Synthese **195** (2018), pp. 4419–4439.

[24] Wright, G. H. V., *Deontic logic*, Mind **60** (1951), pp. 1–15.

A Numeric Default Reasoning System as a Framework for Ethical AI

Joris Graff [1]

University of Utrecht, Department of Philosophy and Religious Studies
Janskerkhof 13, Utrecht, The Netherlands

Abstract

Machine ethics aims to produce moral behaviour in artificial intelligence (AI) systems, by equipping such systems with moral reasoning capacities. One approach within machine ethics is to use deontic logics, i.e. logics that model moral reasoning. Horty has proposed a particular branch of default logic as a potential basis for deontic logic. Default logic models many intuitively desirable features of moral reasoning, such as the possibility of moral conflicts, the idea that moral rules can be overridden, and the distinction between different types of moral reasons. However, when considered in the context of AI ethics, Horty's approach has some shortcomings. Specifically, the traditional, binary valuation of propositions appears unable to capture realistic decision-making scenarios, in which moral reasons can have a variety of strengths. Therefore, this paper explores a numeric default reasoning system, which extends Horty's default logic with numeric valuations for propositions and default rules. It is shown that this new system preserves several advantages of Horty's logic, but can also uniquely model certain intuitively common patterns of moral reasoning, specifically aggregation of reasons. Further directions are suggested, including exploiting the neurosymbolic character of the reasoning system to facilitate moral learning.

Keywords: machine ethics, default logic, moral reasons, aggregation

1 Introduction

As autonomous artificial intelligence (AI) systems become more prominent in multiple societal domains, the question arises how we can program such systems to behave morally. One approach is to equip AI systems with moral rules that, when applied to decision-making situations, lead to morally right decisions with some reliability. This approach is known as *machine ethics* ([3,1,36,2,8]). One approach within machine ethics is to use *deontic logics*, i.e. logics that model moral reasoning and, as such, can be used to derive morally correct decision from representations of decision-making situations in a formal way (see e.g. [35,4,6]). If these derivations can be implemented computationally, such logics

[1] j.j.graff@uu.nl

lend themselves to being incorporated in artificial systems to enable them to make morally right decisions.

In this paper, we consider logic-based machine ethics from a numeric default logic perspective. Default logic is a branch of non-monotonic logic that lends itself well to modelling defeasible reasons. Section 2 outlines the motivation for approaching machine ethics from the perspective of default logic. Section 3 discusses a promising approach to modelling moral reasoning by default logic advanced by John Horty, but section 4 notes some limitations of this approach. Section 5 proposes a numeric extension of Horty's default logic, that associates propositions and reasoning steps with numeric strengths. Section 6 shows that the proposed numeric default logic can reconstruct a limited but core version of Horty's default logic, and section 7 shows that it has additional advantages. Section 8 discusses related literature and section 9 suggests further directions in which numeric default logic may be developed, including facilitating moral learning. Finally, section 10 discusses some philosophical questions raised by numeric default logic and concludes.

Because of space constraints, the system proposed here is only sketched in a rough and exploratory way. Several important patterns of moral reasoning are left to future extensions of the framework, and a deeper exploration of meta-theoretic properties is also left for future work (see also section 9). The main purpose of this paper is to suggest that non-numeric default logic, although capturing some intuitive aspects of moral reasoning, falls short in other respects, and that numeric default logic is a promising avenue to overcome these problems.

2 Motivation

Before considering a specific deontic logic as a framework for machine ethics, we need to know which features of moral reasoning we want the logic to model. On a general level, it is important that deontic logics for ethical AI are a) able to recommend the morally correct outcome in a range of situations and b) explainable to users or other affected parties. To meet the latter requirement, it is preferable to have a logic that accommodates some core features of commonsense moral reasoning. The following four features appear especially important to incorporate in machine ethics:

1) Commonsense moral reasoning is *pluralistic* (see e.g. [18]) That is, there are multiple moral values (e.g. harm minimisation, honesty, respect for autonomy, etc.) giving rise to multiple moral principles. This feature is important for machine ethics, since an AI system that follows a monistic ethical theory, such as Kantianism or utilitarianism, would behave in ways that most people would consider morally problematic, because most people are not strict Kantians or utilitarians. Moreover, a pluralistic ethical framework can more easily accommodate a compromise between different moral concerns that different stakeholders find important.

2) Commonsense moral reasoning is *defeasible*. That is, (most) moral rules

or principles do not hold unconditionally, but can be defeated by other principles. If we accept feature 1, it is likely that in some cases, two different moral principles recommend different actions – for instance, keeping our promise may cause harm. Such cases may be called *moral conflicts.* In moral conflicts, one of both principles is intuitively overridden, and which principle is overridden plausibly depends on the context (e.g. the seriousness of the promise or the severity of harm). Ethical AI systems should incorporate this intuition to reach intuitively correct outcomes in a wide range of cases.

3) Commonsense moral reasoning gives *some* role to aggregation of effects. Intuitively, there are situations in which the duties to prevent effects A and B do not individually outweigh some other duty, but, in combination, do outweigh this other duty. For instance, it may be that I am not released from my promise if keeping it would disappoint friend A, or if keeping it would disappoint friend B, but that I *am* released from my promise if keeping it would disappoint both friends. Ethical AI should incorporate this feature, since otherwise, whenever a moral duty outweighs any of a potentially large set of harms, it also outweighs all those harms together. If implemented, this would lead to an artificial system that would choose to cause major harms.

4) In commonsense moral reasoning, there are different ways in which moral reasons may impact which action ought to be done ([25]). Jonathan Dancy ([10], pp. 42-43) has made the following useful distinction between three types of moral reasons [2] :
   - A *(dis)favourer* is simply a reason that counts in favour of or against an action.
   - An *intensifier* is a reason that does not directly count in favour of or against an action, but rather increases the strength of another reason. Conversely, an *attenuator* decreases the strength of another reason.
   - A *disabler* is a reason that does not directly count in favour of or against an action, but rather removes another reason from consideration (as opposed to merely reducing its strength).

   A short set of examples will illustrate the plausibility of this distinction. Imagine first that I promised to perform act A, but that doing A would result in some harm. Then the fact that I promised to do A (call this fact P) serves as a favourer of doing A, and the fact that A would cause harm (call this fact H) serves as a disfavourer of doing A. Imagine, second, that P and H are again the case, but that the person to whom I promised to do A is a close friend who has shown much loyalty to me in the past. Plausibly, this does not itself give me a reason to do A, but it does increase the strength of my promise; i.e. it serves as an intensifier of P. Imagine, third, that P and H are again the case, but that the promisee

[2] We here depart from Dancy's terminology, in which only favourers and disfavourers are called reasons.

has explicitly released me from the promise. Then the promise no longer has *any* strength; hence this fact serves as a disabler of P. Whether of not P defeats H as a reason may depend both on the initial strengths of P and H and on the effects of potential intensifiers, attenuators or disablers.

One question about attenuators and intensifiers is whether they decrease or increase the strength of a reason *comparatively* to another reason, or whether they decrease or increase the strength of a reason in general (following Horty ([23], p. 145), we will call this *per se*). In the first case, an intensifier, for instance, would be a reason for the conclusion 'reason A is more important than reason B', whereas in the second case, the conclusion would be of the form 'A is more important (full stop)'. The difference becomes clear when we want to compare A to a third reason, say C. The comparative intensifier tells us nothing about how A compares to C. However, the *per se* intensifier may lead us to conclude that, since A is stronger in general, it is now also stronger than C (provided that A and C were equally strong to begin with). Arguments can be made for both interpretations, or for a distinction between two types of intensifiers (and attenuators). In this paper, we will stick to a *per se* interpretation (as does Dancy in [10]). The reason is that, in cases of moral intensification and attenuation, we intuitively have grounds to increase or diminish the strength of a reason A without having to specify each individual other reason in comparison to which A's strength is increased or diminished. For instance, plausibly, the fact that someone is a close friend serves as an intensifier of reasons concerning that person's suffering – i.e., her suffering counts more. We do not have to specify a range of individuals in comparison to whom our friend's suffering counts more (indeed, such a specification will always fall short). More generally, given that a reason A can always come to clash with other, perhaps unforeseen reasons, it seems that specifying a limited set of reasons in comparison to which A's strength is intensified leaves open many potential comparisons. But in many cases, this comparison ought to be settled by the fact that one reason is intensified (or attenuated) and the other is not.

A moral system that incorporates the first three features is Ross's theory of *prima facie* duties ([33]). A *prima facie* duty is a feature of a moral act that defeasibly tends to make the act morally right or wrong. In Ross's theory, moral agents have multiple types of duties, such as the duty to keep our promises and the duty to prevent harm. When different duties conflict, one of the duties can be overridden. When an action involves positive or negative consequences, whether or not the duty to prevent or ensure those consequences is overridden may depend on the aggregation of the consequences. Ross's framework can be extended to include the fourth feature, i.e. the distinction between different types of moral reasons, by allowing that the strength of *prima facie* duties, which in specific situations serve as (dis)favourers, is often modified, or entirely removed, by further considerations, which serve as intensifiers, attenuators and disablers.

In the next section, we turn to Horty's brand of default logic as a promising candidate for formally modelling this extended Rossian framework. This choice is motivated by the fact that Horty explicitly intends his framework to be a model of different types of reasons, along the lines outlined above ([23], pp. 41–61, 147–165), and also by the fact that default logic is inherently defeasible and therefore naturally allows for modelling moral conflicts. This latter feature may be seen as an advantage over modal deontic logics, which are generally not defeasible and therefore struggle with representing moral conflicts. [3]

3 Horty's Default Logic

'Default logic' refers to a class of non-monotonic logics originally introduced by Reiter ([32]). We are here interested in Horty's default logic (HDL), which only contains what Reiter calls 'normal defaults'. This section summarises the main elements of HDL, as set out in [23], specifically Horty's *variable default theories.*

HDL is based on a language $\mathcal{L}$ which is some ordinary logical language, e.g. propositional logic, enriched with constants of the form d_x, where each such constant refers to a default δ_x, a relation symbol $\prec$, such that $d_x \prec d_y$ means that δ_y has a higher priority than δ_x (written as $\delta_x < \delta_y$), and a predicate *Out*, such that $Out(d_x)$ means that δ_x is disabled. Then, a (variable) default theory Δ is defined as follows:

Definition 3.1 A *default theory* Δ is an ordered pair $\langle \mathcal{W}, \mathcal{D} \rangle$, where $\mathcal{W}$, the *knowledge base*, is a set of propositions in $\mathcal{L}$ and $\mathcal{D}$ is a set of *defaults* of the form $\delta : X \to Y$, where X and Y, the default's *premise* and *conclusion* respectively, are again propositions in $\mathcal{L}$.

We write $prem(\delta) = X$ and $conc(\delta) = Y$. We also write $prem(\mathcal{S})$ and $conc(\mathcal{S})$ to denote the union of all premises or conclusions of a set $\mathcal{S}$ of defaults. It is imposed that $\mathcal{L}$ contains a constant d_x for each $\delta_x \in \mathcal{D}$. Intuitively, we can think of a default theory Δ as a set of known statements ($\mathcal{W}$) and a set of defeasible rules ($\mathcal{D}$) that can be used to extract conclusions from those statements.

These extracted conclusions are formalised as *extensions*, which in turn are built on the concept of *scenarios*. A scenario $\mathcal{S}$ is a subset of $\mathcal{D}$, intuitively consisting of those defaults that have been applied to the knowledge base. A *proper* scenario is then defined as follows.

Definition 3.2 Given a default theory $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$, a *proper* scenario $\mathcal{S} \subseteq \mathcal{D}$ is a scenario that contains all and only those defaults that are a) *triggered*, b) *not conflicted* and c) *not defeated* in the context of that scenario.

[3] This is a slight oversimplification. Horty ([22]) shows that modal deontic logics can, to some extent, accommodate moral conflicts. However, it is unclear how they can model resolution of these conflicts by higher-order reasons (attenuators etc.), which play an important role in the framework set out here. See also Fuhrmann ([15]) for further discussion and Dong et al. ([11]) for a proposal to embed modal deontic logic in a non-monotonic framework.

- A default δ is *triggered* in the context of scenario $\mathcal{S}$ iff a) $\mathcal{W} \cup \mathit{conc}(\mathcal{S}) \vdash \mathit{prem}(\delta)$ and b) δ is not *disabled*, i.e. $\mathcal{W} \cup \mathit{conc}(\mathcal{S}) \nvdash \mathit{Out}(d)$.
- A default δ is *conflicted* in the context of scenario $\mathcal{S}$ iff $\mathcal{W} \cup \mathit{conc}(\mathcal{S}) \vdash \neg \mathit{conc}(\delta)$.
- A default δ is *defeated* in the context of scenario $\mathcal{S}$ iff there is a default δ' which is triggered in the context of $\mathcal{S}$, such that a) $\mathcal{W} \cup \mathit{conc}(\mathcal{S}) \vdash d \prec d'$ and b) $\mathcal{W} \cup \{\mathit{conc}(\delta')\} \vdash \neg \mathit{conc}(\delta)$.

Intuitively, a proper scenario contains all defaults that can consistently be applied because their premises are known and their conclusions are not known to be false on the basis of other defaults in the scenario or of stronger, applicable defaults outside of the scenario. Then an extension $\mathcal{E}$ is defined as follows:

Definition 3.3 Given a default theory $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$ and a proper scenario $\mathcal{S}$ of Δ, an *extension* of Δ is $\mathcal{E} = Th(\mathcal{W} \cup \mathit{conc}(\mathcal{S}))$,

where Th denotes logical closure. Note that a default theory can have multiple proper scenarios, and therefore multiple extensions.

So far, the logic can be applied to any form of defeasible reasoning and is not clearly a deontic logic. Horty ([23], pp. 68–74) suggests extending it into a deontic logic by defining a deontic operator $\bigcirc$ in terms of extensions. Here, we take a different approach: we use default logic to reason *about* moral statements, rather than defining moral statements in terms of default theories (cf. [15]). This means that the logic is not deontic in the classical sense of containing deontic operators, but rather that the logic itself has been tuned to the features of moral reasoning, as discussed in section 2.

Horty ([23], pp. 42–47) argues that, given a default theory, a *reason* for some proposition X can be defined as a proposition that is the premise of a triggered default that has X as its conclusion. Moreover, we can, within HDL, distinguish between the three types of reasons identified in the previous section ([23], pp. 142–146). First, a *favourer* of a proposition X is just a proposition that is the premise of a triggered default that has X as its conclusion. Second, when we have a triggered default of the form $\delta_1 : X \rightarrow d_2 \prec d_3$, then X is an *intensifier* of δ_3 and an *attenuator* of δ_2.[4] Third, when we have a triggered default of the form $\delta_1 : X \rightarrow \mathit{Out}(d_2)$, then X is a *disabler* of δ_2.

Example 3.4 *As an example of the application of HDL to machine ethics, consider the following situation that a self-driving car may encounter. Imagine that, suddenly, two pedestrians appear in front of the car, so that there is no option to brake before the pedestrians are hit and thereby bodily harmed. The*

[4] Note that in this definition, contrary to the definition given in section 2, attenuators and intensifiers are *comparative*, i.e. they always attenuate or intensify some reason compared to some other reason. As Horty notes, this feature is 'dictated by our framework assumption that reasons are to be related to one another only through an ordinal ranking, rather than an assignment to each of some cardinal weight, so that strengthening one reason relative to another is the only for of strengthening there is' ([23], p. 145). But as we saw in section 2, this notion of attenuation has some counter-intuitive results. We will return to this issue in section 6, after we have introduced a cardinal weighing of reasons.

only way to avoid hitting the pedestrians is to steer the car away from them. This would cause it to crash into the guardrail, causing bodily harm to the passenger. Imagine, moreover, that both pedestrians appear in front of the car as the result of a demonstrable traffic violation (e.g. ignoring a red traffic light). We can model this situation as a default theory $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$. Let G be the proposition that steering into the guardrail would harm the passenger, P_2 the proposition that not steering would harm the two pedestrians in front of the car, V the proposition that the pedestrians violated a traffic rule, and S the proposition that the car ought to steer to the side. Then, $\mathcal{W} = \{G, P_2, V\}$. Clearly, the default set $\mathcal{D}$ should contain the defaults $\delta_1 : G \rightarrow \neg S$ and $\delta_2 : P_2 \rightarrow S$, since there is a general prima facie duty to prevent harm to both the passenger and the pedestrians. Moreover, the fact that there are two pedestrians as opposed to one passenger plausibly gives rise to $\delta_3 : \top \rightarrow d_1 \prec d_2$, which gives a general reason to prioritise the pedestrians, following from the vacuously true proposition $\top$. Finally, the fact that the pedestrians are in their current predicament due to violating a traffic rule may be thought to give rise to $\delta_4 : V \rightarrow d_2 \prec d_1$, giving a reason to prioritise the innocent passenger.[5] *This default theory can be displayed in graph form as in Fig. 1, adopting Horty's conventions that $X \Rightarrow Y$ means that Y is a strict logical consequence of X, $X \rightarrow Y$ means that Y follows from X by default, and $X \not\Rightarrow Y$ and $X \not\rightarrow Y$ mean the same as $X \Rightarrow \neg Y$ and $X \rightarrow \neg Y$, respectively. $\top \Rightarrow X$ means that X follows from the trivially true proposition, which is a way to indicate that X is in $\mathcal{W}$.*

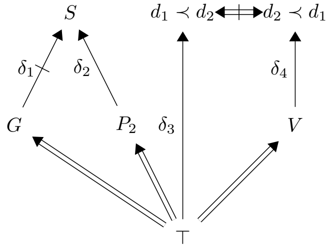


Fig. 1. Example III.1.

In this default theory, we have two types of reasons. The fact that the passenger would be harmed by steering into the guardrail is a favourer for not steering (or a disfavourer for steering), and the fact that two pedestrians would be harmed otherwise is a favourer for steering. At the same time, the trivial proposition acts as a general intensifier of the potential harm to the pedestrians

[5] This latter default is of course controversial. The point is not that the reader should agree with it, but only to illustrate the functioning of attenuators.

as a reason to steer, while the fact that the pedestrians are guilty of a traffic violation serves as an attenuator of the reason provided by their potential harm.

To determine which action recommendation can be inferred from this moral default theory, we need to find its extensions. Δ has two proper scenarios: $\mathcal{S}_1 = \{\delta_1, \delta_4\}$ and $\mathcal{S}_2 = \{\delta_2, \delta_3\}$ (the reader can verify that these are the only scenarios satisfying the criteria in definition 3.2). Therefore, there are two extensions: $\mathcal{E}_1 = Th(\{G, P_2, V, \neg S, d_2 \prec d_1\})$, and $\mathcal{E}_2 = Th(\{G, P_2, V, S, d_1 \prec d_2\})$. Since we are faced with multiple, incompatible extensions, we are faced with a dilemma whether or not to steer. In order to force a decision, we would need to provide a priority ordering between δ_3 and δ_4. For instance, we may add the sentence $d_3 \prec d_4$ to $\mathcal{W}$. This new default theory would have only one extension, containing $\neg S$, the recommendation not to steer.

4 Limitations of Horty's Default Logic

HDL satisfies features 1, 2 and 4 outlined in section 1: 1) it can accommodate multiple duties, modelled as different defaults, 2) its rules are inherently defeasible, and 4) it can model different types of reasons, as seen in the previous section. However, as Horty ([23], pp. 59–61) notes, it does not satisfy feature 3: it does not allow for aggregation of reasons. That is, if a default theory Δ contains (in $\mathcal{D}$) a set $\mathcal{Z}$ of defaults for some conclusion Z, and each default in this set is individually outweighed by some other default δ_k for the opposite conclusion $\neg Z$, then the set as a whole is also outweighed by default δ_k. This means that, unless Z follows from Δ in some other way, or unless δ_k is disabled, no extension of Δ will contain Z, no matter how many members of $\mathcal{Z}$ are triggered.

To return to the example in section 2: say that I promised to do A (call this reason P), but that doing A would cause harm to two of my friends (call these reasons F_1 and F_2). This situation can be modelled as a default theory $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$, with $\mathcal{W}$ containing P, F_1, and F_2, and with $\mathcal{D}$ containing $\delta_1 : P \to A$, $\delta_2 : F_1 \to \neg A$, and $\delta_3 : F_2 \to \neg A$. Say moreover that the promise outweighs both harms individually, i.e. $\mathcal{W}$ contains both $d_2 \prec d_1$ and $d_3 \prec d_1$. Then, if there are no other sentences in $\mathcal{W}$ and no other defaults in $\mathcal{D}$, Δ's only extension will contain A. But this may be counterintuitive: it may be that both harms, though each is individually outweighed by P, *together* ought to outweigh P. This situation may be mended by adding a fourth default, $\delta_4 : F_1 \wedge F_2 \to \neg A$, and adding the sentence $d_1 \prec d_4$ to $\mathcal{W}$. But this solution is unsatisfactory. If there would be many harms that are individually outweighed by P, but that, in various combinations, outweigh P, we would need to add a separate default for each combination of harms that could possibly outweigh P. This would make the default theory very cumbersome and difficult to interpret. Moreover, it would make it difficult to update the default theory when new defaults become known.

In recent years, some proposals have been made to model aggregation (or more generally 'accrual' – this notion is more general in that accruing reasons may weaken as well as strengthen each other) in defeasible reasoning systems.

Prakken's ASPIC$^+$-based approach ([31]) overcomes the problem of having to compare many sets of accruing reasons (or, in ASPIC$^+$, arguments) by only comparing the sets of those accruing arguments ('accrual sets') that are in fact active in a specific situation (more specifically, given an in/out-labelling, only arguments are considered that are not labelled 'out' and that have no undercutter labelled 'in'). However, in different situations, this method may still produce many different accrual sets. In a machine ethics context, this means that we may still need to provide an unfeasibly large number of ordinal rankings between potential accrual sets to deal with all situations that an AI system may encounter. The alternative would be to only provide rankings *after* the accrual sets are established, i.e. after it is known which arguments are and are not active. But this is also unfeasible, since it would mean that, for any decision-making situation that the AI system faces, a weighing of reasons can be provided only after the details of the situation are known. This would require continuous oversight by a human supervisor, in effect abandoning the field of machine ethics and returning moral decision-making to the supervisor. It would be much preferable if a default theory that contains many weak reasons would automatically aggregate these reasons without requiring an explicit weighing. In order for this to be possible, however, we need to associate each default with a certain numeric *weight*. This approach is taken up in Gordon and Walton's Carneades-based approach ([20]), further developed by Gordon ([19]). This approach introduces *weighing functions* that assign a weight between 0 and 1 to each argument within a labelled (i.e. valuated) argument graph, on the basis of the in/out-labelling and further details of the graph, which may include external information about the plausibility of arguments. Aggregation may then be modelled by positing a weighing function that makes an argument's weights a positive (partial) function of the number and strength of its premises. This approach is promising in the current context, but the structure of the argument graphs it is based on does not allow for a distinction between all types of reasons set out in section 2 (specifically, it does not account for attenuators and intensifiers). Below, we therefore develop an approach that combines numeric valuations with the ability to distinguish between different types of reasons.

There is a second, related problem with HDL as a basis for machine ethics. In Horty's system, knowledge is represented as propositions with discrete valuations. But the moral domain that some autonomous systems would likely encounter is too complex to be captured in this way. The reason is that some morally relevant features, such as amount or probability of harm, allow for many gradations. For instance, to capture each potential situation that a self-driving car may encounter, we would need propositions such as 'Probability of 0.2–0.25 of 1 non-passenger accruing harm with severity in range 0.5–0.55 by action 1'. This would lead to a very cumbersome language. It would be preferable to have propositions with continuous valuations. For instance, the strength of the proposition 'Harm may occur to person 1' may vary given the probability and severity of the harm to person 1. This requires a system that assigns weights, not only to defaults, but also to propositions.

5 Numeric Default Logic

The aim of this section is to develop a *numeric default logic* (NDL)[6] that preserves some features of HDL, most notably its ability to model different types of reasons, but that introduces numeric weights for both propositions and defaults in order to overcome the limitations mentioned in section 4.

We define NDL by replacing default theories with default *graphs*.

Definition 5.1 A *default graph* Γ is a triple $\langle \mathcal{N}_{\mathcal{P}}, \mathcal{N}_{\mathcal{D}}, \mathcal{C} \rangle$, where $\mathcal{N}_{\mathcal{P}}$ is a set of *proposition nodes*, $\mathcal{N}_{\mathcal{D}}$ is a set of *default nodes*, and $\mathcal{C} \subseteq \{\langle n_1, n_2 \rangle | n_1, n_2 \in \mathcal{N}_{\mathcal{P}} \cup \mathcal{N}_{\mathcal{D}}\}$ is a set of *connections* or *edges* between nodes. (Thus, given $\mathcal{N} = \mathcal{N}_{\mathcal{P}} \cup \mathcal{N}_{\mathcal{D}}$, the pair $\langle \mathcal{N}, \mathcal{C} \rangle$ is a *directed graph.*)

We sometimes write c_{ab} as a shorthand for the connection $\langle n_a, n_b \rangle$. When, for two nodes n_1 and n_2, there is a connection $\langle n_1, n_2 \rangle$, we say that n_1 is an input node of n_2 and n_2 is an output node of n_1. We also say that n_1 *feeds into* n_2.

Default graphs are subject to the following constraint: for each default node $n_\delta \in \mathcal{N}_{\mathcal{D}}$, there is at least one node $n \in \mathcal{N}$ such that $\langle n, n_\delta \rangle \in \mathcal{C}$ and there is exactly one node $n \in \mathcal{N}$ such that $\langle n_\delta, n \rangle \in \mathcal{C}$. Intuitively, nodes that feed into a default node represent either the premise of the corresponding default or other defaults that affect the weight of the default. The node that a default node feeds into represents the conclusion of the corresponding default.

Furthermore, we call the set of nodes $\mathcal{N}_{in} = \{n \in \mathcal{N} | \neg \exists n' \in \mathcal{N} : \langle n', n \rangle \in \mathcal{C}\}$ the set of *input nodes* and we call the set of nodes $\mathcal{N}_{out} = \{n \in \mathcal{N} | \neg \exists n' \in \mathcal{C} : \langle n, n' \rangle \in \mathcal{C}\}$ the set of *output nodes.* Note that the above constraint guarantees that $\mathcal{N}_{in} \subseteq \mathcal{N}_{\mathcal{P}}$ and $\mathcal{N}_{out} \subseteq \mathcal{N}_{\mathcal{P}}$, i.e. only proposition nodes can be input or output nodes.

Next, the question is how to derive conclusions from premises in default graphs. In this paper, we limit ourselves to *acyclic* default graphs, i.e. default graphs such that $\langle \mathcal{N}, \mathcal{C} \rangle$ forms a directed acyclic graph (DAG). We also limit ourselves to proposition nodes that correspond to *atomic* propositions; that is, we do not consider connectives such as conjunction and disjunction. Given these limitations, the strategy for derivation within default graphs is as follows. We first introduce a *valuation* $\mathcal{V} : \mathcal{N} \cup \mathcal{C} \to \mathbb{R}$, which is a function that assigns a numerical value to each node and to each connection. $\mathcal{V}$ is subject to the constraint that the value of each non-input node is a function of the values of its input nodes and the corresponding connections. More precisely, for each non-input node n_k, we define a numeric *activation function* g_k. Then values of non-input nodes are determined as follows:

[6] 'Logic' is here used in the wide sense of a formal reasoning system. NDL is not a logic in the more restricted sense, since it does not define a formal proof system. We use the term 'numeric default logic' rather than 'numeric default reasoning system' to stress the affinity with HDL.

- For each proposition node $n_P \in \mathcal{N}_\mathcal{P}$ with input nodes n_i,

$$V(n_P) = g_P(\sum_i V(n_i)V(c_{iP})) \tag{1}$$

  where $V(n_i)$ is the valuation of input node n_i and $V(c_{iP})$ is the valuation of the connection $\langle n_i, n_P\rangle$.

- For each default node $n_\delta \in \mathcal{N}_\mathcal{D}$ with input nodes n_i,

$$V(n_\delta) = g_\delta(\prod_i V(n_i)V(c_{i\delta})) \tag{2}$$

  where $V(n_i)$ is the valuation of input node n_i and $V(c_{i\delta})$ is the valuation of the connection $\langle n_i, n_\delta\rangle$.

The difference between functions 1 and 2 has the following motivation. Different inputs to a proposition node n_P represent different favourers or disfavourers of the conclusion P. Favourers and disfavourers appear to function in an additive manner; hence the activation of n_P is a function of the *sum* of its inputs. On the other hand, different inputs to a default node n_δ represent different factors that impact the strength of the default. These include the default's premise, but also attenuators, intensifiers, or disablers of the default in question. These last three types of reasons do not directly add to or subtract from a default's strength, but rather strengthen or lessen it proportionally. Intuitively, an attenuator multiplies a default's strength by a value lower than 1 (but higher than 0), an intensifier multiplies a default's strength by a value higher than 1, and a defeater multiplies a default's strength by 0 (see also [25]). Therefore, the activation of n_δ is a function of the *product* of its inputs.

Next, we introduce a *positive threshold value* θ_p and a *negative threshold value* θ_n. Given these elements, we define a *valuated default graph* as follows:

Definition 5.2 Given a default graph $\Gamma = \langle \mathcal{N}_\mathcal{P}, \mathcal{N}_\mathcal{D}, \mathcal{C}\rangle$, a *valuated default graph* Γ_V on Γ is defined as a tuple $\langle \Gamma, V, \mathcal{G}, \theta_p, \theta_n\rangle$, where $V : \mathcal{N} \cup \mathcal{C} \to \mathbb{R}$ is a valuation function, $\mathcal{G} = \{g_k : \mathbb{R} \to \mathbb{R} | n_k \in \mathcal{N} \setminus \mathcal{N}_{in}\}$ is a set of activation functions g_k for each non–input node n_k, and $\theta_p, \theta_n \in \mathbb{R}$ are a positive and a negative threshold.

Given a valuated default graph, and an interpretation according to which each proposition node n_P corresponds to a proposition P, we can define extensions as follows:

Definition 5.3 Given a valuated default graph $\Gamma_V = \langle \Gamma, V, \mathcal{G}, \theta_p, \theta_n\rangle$, an extension $\mathcal{E}$ is defined as $\mathcal{E} = Th(\{P | V^{-g}(n_P) > \theta_p\} \cup \{\neg P | V^{-g}(n_P) < \theta_n\})$. Here, V^{-g} is the activation of a node *before* the activation function g has been applied, i.e.

- For each non–input proposition node $n_P \in \mathcal{N}_\mathcal{P} \setminus \mathcal{N}_{in}$ with input nodes n_i,

$$V^{-g}(n_P) = \sum_i V(n_i)V(c_{iP}) \tag{3}$$

- For each default node $n_\delta \in \mathcal{N_D}$ with input nodes n_i,
$$V^{-g}(n_\delta) = \prod_i V(n_i)V(c_{i\delta}) \tag{4}$$
- For each input proposition node $n_P \in \mathcal{N_P} \cap \mathcal{N}_{in}$, $V^{-g}(n_P) = V(n_P)$.

Obviously, each valuated default graph has a single extension.

This general framework allows, of course, for an indefinitely large variety of interactions between reasons, since there are indefinitely many activation functions we could choose for each non-input node. The next section will show, however, that, using a limited number of activation functions, we can reconstruct Horty's default theories in an intuitive manner.

6 Reconstructing Horty's Default Logic in NDL

NDL can reconstruct a restricted version of HDL, in that we only consider default theories Δ that meet the following conditions. (Below, we call atomic propositions of the form P, that do not contain the $\prec$ or *Out* operators, *normal* atomic propositions. Similarly, we call literals of the form P or $\neg P$ *normal* literals).

(i) Propositions in $\mathcal{W}$ are restricted to a) normal literals, b) sentences of the form $d_1 \prec d_2$ or $Out(d_1)$, and c) implications of the form $P \supset \Phi$, where P is a normal atomic proposition and Φ is a normal literal.

(ii) $\mathcal{W}$ is consistent.

(iii) Defaults in $\mathcal{D}$ are restricted to defaults where the premise is a normal atomic proposition P and the conclusion is a normal literal or a sentence of the form $d_1 \prec d_2$ or $Out(d_1)$.

(iv) $\mathcal{D}$ is finite.

(v) Δ is acyclic. That is to say, there is no series $Q_1, P_1, \ldots, Q_{k-1}, P_{k-1}, Q_k$ such that $Q_1 = Q_k$ and such that for each $i, 1 \leq i < k$, a) either $Q_i = P_i$ or $Q_i = \neg P_i$ or $\mathcal{D}$ contains a default of the form $\delta : P_i \rightarrow Q_{i+1}$ and $Q_i = Out(d)$ or $Q_i = d \prec d'$ or $Q_i = d' \prec d$ for some $\delta' \in \mathcal{D}$, and b) either there is some $\delta : P_i \rightarrow Q_{i+1} \in \mathcal{D}$ or there is some $\delta : P_i \rightarrow \neg Q_{i+1} \in \mathcal{D}$ or $\mathcal{W} \cup \{P_i\} \vdash Q_{i+1}$ or $\mathcal{W} \cup \{P_i\} \vdash \neg Q_{i+1}$.

These restrictions hold for most examples that Horty discusses in [23], with two exceptions. First, Horty sometimes discusses default theories where $\mathcal{W}$ or the premises or conclusions of $\mathcal{D}$ contain conjunctions or disjunctions. However, such complex sentences are in effect treated the same as atomic sentences and therefore, in our restricted language, can be replaced by introducing a new atomic sentence. Second, Horty discusses some cases in which $\mathcal{W}$ or the conclusions of $\mathcal{D}$ contain sentences of the form $\neg(A \wedge B)$, indicating that two atomic propositions are mutually exclusive. Such sentences would have to be modelled by cyclic default graphs, which belong to future work (see section 9).

Following the strategy of D'Avila Garcez et al. ([17], ch. 3, [16], ch. 4) for representing defeasible logic programs as neural networks, we have first constructed an algorithm to translate each default theory Δ in HDL into a default graph Γ in NDL, and to translate each extension $\mathcal{E}$ of Δ into a valuated default graph Γ_V based on Γ. Next, we have proven that extensions in HDL can be reconstructed by default graphs as obtained by this algorithm. Because of space constraints, the details of the translation algorithm and the proof are omitted. Instead, we provide the core principles behind the translation, the precise result that was proven, and an example.

- Normal atomic propositions in HDL correspond to proposition nodes in NDL. Each proposition node has the step function $step(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$ as its activation function.
- Defaults in HDL correspond to default nodes in NDL. For each default δ with premise P, the proposition node n_P feeds into the default node n_δ in NDL (recall that we only consider defaults with atomic propositions as premises). There are three types of default nodes:
  - · If a node n_δ represents a favourer or disfavourer default (i.e. a default of the form $\delta : P \to Q$ or $\delta : P \to \neg Q$, with P and Q being normal atomic propositions) in HDL, then n_δ feeds into n_Q. Since, intuitively, a more strongly activated (dis)favourer provides a stronger reason, n_δ has the linear function $lin(x) = x$ as its activation function.
  - · If a node n_δ represents an attenuator default (i.e. a default of the form $\delta : P \to d' \prec d''$) in HDL, then n_δ feeds into the node corresponding to the default that is attenuated, i.e. node $n_{\delta'}$. The activation function should be such that, if activated, node n_δ multiplies the activation of node $n_{\delta'}$ by a value between 0 and 1. Below, we consider the function $\tau(x) = -tanh(x) + 1$, which satisfies this property.
  - · If a node n_δ represents a disabler default (i.e. a default of the form $\delta : P \to Out(d')$) in HDL, then n_δ feeds into the node corresponding to the default that is disabled, i.e. node $n_{\delta'}$. The activation function should be such that, if activated, node n_δ multiplies the activation of node $n_{\delta'}$ by a value of 0. Below, we consider the inverse step function $invstep(x) = \begin{cases} 1 & \text{if } x \leq 0 \\ 0 & \text{if } x > 0 \end{cases}$.
- Material implications in HDL correspond to direct connections from the antecedent to the consequent nodes.

This translation is not complete; specifically, it models attenuators differently than HDL. The reason is that NDL involves a conceptual shift in the interpretation of attenuators (and intensifiers) away from Horty's comparative interpretation and towards Dancy's *per se* interpretation. That is, whereas in HDL attenuators always compare two defaults, in NDL, each attenuator node outputs to exactly one default node. As argued in section 2, this conceptual shift is intuitive, but it does mean that NDL cannot represent sentences of the form $d_1 \prec d_2$, which HDL can represent (see also footnote 4). Despite this difference, the (partial) translation produced by the algorithm based on the above translation principles still reconstructs HDL's extensions with regards to normal literals. That is, we have proven the following correspondence: for a default theory Δ in HDL, and an extension $\mathcal{E}$ of Δ, if we construct a valuated default graph Γ_V according to the above translation principles, then Γ_V's extension $\mathcal{E}'$ contains exactly the same normal literals as does $\mathcal{E}$. Note that this result is obtained using only a small number of activation functions that intuitively capture the roles of favourers, disfavourers, attenuators, and disablers. Therefore, the result shows not only that it is possible to reproduce extensions in HDL in NDL (which in itself would be unsurprising, given the freedom to choose different activation functions), but also that this is possible in a natural way.

6.1 Example

To illustrate the reconstruction of HDL, we show a valuated default graph that reconstructs an extension of the default theory from example 3.1.

Example 6.1 *As we saw in example 3.1, Δ has two extensions. The valuated default graph shown in Fig. 2 reconstructs extension $\mathcal{E}_1$, which is derived from the scenario $S_1 = \{\delta_1, \delta_4\}$. (Extension $\mathcal{E}_2$ can be reconstructed by providing different valuations on the same default graph.) Proposition nodes are indicated by squares and default nodes are indicated by circles. Valuations are given to the right of each node or connection. For nodes, the activation function g and the pre-activation value V^{-g} are also given.*

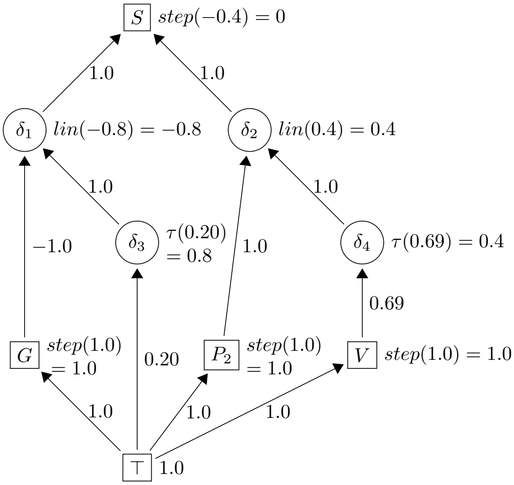


Fig. 2. Example VI.1.

We can see that this default graph indeed yields an extension $\mathcal{E}_1'$ such that for all normal propositions P, $P \in \mathcal{E}_1'$ iff $P \in \mathcal{E}_1$ and $\neg P \in \mathcal{E}_1'$ iff $\neg P \in \mathcal{E}_1$. Moreover, the graph captures the idea that, in scenario $\mathcal{S}_1$, δ_4 outweighs δ_3 and therefore δ_1 outweighs δ_2: n_{δ_4} attenuates the activation of n_{δ_2} more than n_{δ_3} attenuates the activation of n_{δ_1}. This makes it possible to interpret the default graph in terms of reasons, as is the case for default theories: the self-driving car should not steer because the guardrail would harm the passenger (a favourer), and because the fact that the pedestrians have committed a traffic violation weakens the reason that their harm provides for steering (an attenuator).

7 Extending Horty's Default Logic

Section 6 has shown that we can reconstruct (a restricted but core version of) HDL in NDL. However, the main interest of NDL is that it can extend HDL, specifically in a way that solves the problems described in section 4. It is clear how we can solve the first problem, i.e. that HDL does not allow for aggregation of reasons. Since a proposition node's valuation is a function of the sum of the valuations of its input nodes, several weaker input nodes could together outweigh a strong input node.

The second problem was that binary valuations cannot capture the complexity of realistic situations. Consider first proposition nodes that are part of the knowledge base, e.g. nodes representing facts such as 'Person 1 may be harmed by action A'. In realistic situations, such nodes should not obtain their input from the fixed node $n_\top$, but rather from a more distributed representation of the moral situation. This representation may for instance be a *vector* representing different circumstances relevant to the probability, strength etc. of the fact that the node in question (call it n_P) represents. For instance, n_P could have one input node corresponding to the number of people harmed, one input node corresponding to the probability of harm occurring to each of these people, etc. When properly weighed, this input vector to n_P would produce an appropriate value for $V(n_P)$ which represents a combination of the probability and intensity of the harm. Alternatively, n_P may receive its input from a neural network that has been trained to estimate the probability and extent of harm on the basis of the system's sensor inputs. The affinity between default graphs and neural networks is explored further in section 9.

Consider next proposition nodes that represent the proposition that an action ought (not) to be performed. Intuitively, we sometimes feel that an action is strongly morally required or forbidden, whereas in other situations the balance of reasons is almost equal (see also section 10). To capture this intuition, such evaluative proposition nodes should have a non-binary activation function, e.g. the linear function ($f(x) = x$), or some normalised continuous function.

By using continuous inputs and activation functions, default graphs can represent a wider and more realistic range of situations than Horty's default theories. At the same time, they retain the features that made HDL a promising framework for deontic logic. Specifically, default graphs still distinguish between different types of moral reasons (which are distinguished by different activation functions), and allow for moral conflicts that may be resolved in a variety of ways.

8 Related Work

Apart from default logic – especially Horty's version – and Gordon and Walton's ([20,19]) approach discussed in section 4, NDL has commonalities with other strands of work within defeasible reasoning. The closest of these is the neurosymbolic approach towards logic programming advanced by D'Avila Garcez et al.; this approach is discussed in section 8.1. After this, we discuss another type of defeasible reasoning systems that lends itself to gradual valuation, namely Pollock's inference graphs (8.2). The current section only focuses on *structured* inference or argumentation systems; it should be noted, however, that there has also been much work on gradual valuations within the context of Dung's unstructured argumentation frameworks ([12]) (see e.g. [7,13,27,26,34,24]).

8.1 Neuro-Symbolic Logic Programming

D'Avila-Garcez et al. ([17], ch. 3, [16], ch. 4) have developed the *Connectionist Inductive Learning and Logic Programming System* (CILP). CILP provides neural network implementations of non-monotonic logic programs. Most notably, D'Avila-Garcez et al. discuss *extended logic programs*, which consist of sets of implicative clauses by which a literal (the 'head') follows from a set of literals (the 'body'), where literals in the body may be preceded by a negation-as-failure operator $\sim$. Each such clause is roughly interpreted as stating that the literal in the head can be defeasibly inferred if all normal literals in the body are known to be true, and all literals preceded by the negation-as-failure operator are not known to be false (they thus have a strong similarity with defaults with potential disablers). The semantics of extended logic programs are defined by *answer sets*, which, roughly, contain all and only those conclusions that can consistently be derived from a knowledge base in combination with a set of clauses (these have a strong similarity to extensions in default logic). D'Avila-Garcez et al. have provided algorithms to translate extended logic programs into three-layer artificial neural networks (with the input layer representing all literals occurring in the body of some clause and the output layer representing all literals occurring in the head of some clause). D'Avila-Garcez et al. ([17], ch. 3) prove that, for extended logic programs that are *well-behaved*, i.e. have just one answer set, provided some threshold to determine which nodes are included in the answer set, the neural networks produced by these algorithms yield the same answer sets as the logic programs they are based on. Conversely, a CILP-style neural network can be learnt from training data rather than based on a pre-given logic program, in which case a logic program can be extracted from the trained model.

Besold et al. ([5]) use CILP as a framework for deontic logic. They do so by interpreting extended logic programs in a deontic manner, which is to say that each clause represents a norm, each literal in the body of the clause contains a precondition for the norm to hold, and the literal in the head of the clause represents an obligation. (Technically, Besold et al. depart from input/output logic, not logic programming. But as they show, their input/output models can straightforwardly be translated into extended logic programs.) Thus for such deontic extended logic programs, each answer set can be interpreted as an 'obligation set'. Then, they show that D'Avila-Garcez et al.'s translation algorithm can be used to construct a neural network that computes the obligation set for any well-behaved deontic extended logic program.

The current approach is in some ways inspired by D'Avila Garcez et al.'s approach, specifically in the usage of translation algorithms and thresholds to construct numeric networks on the basis of non-numeric non-monotonic models. D'Avila Garcez et al.'s motivations are very different from the motivations behind NDL, however. Specifically, D'Avila Garcez et al.'s neural networks are meant to correspond directly to extended logic programs (the added value being that they facilitate learning of logic programs), whereas default graphs in NDL are meant to go beyond default theories in HDL. Thus, the correspondence in CILP is two-way: each logic program can be represented as a neural network, and each CILP-style neural network can also be interpreted as a logic program. This is not the case for HDL and NDL: although each default theory (that meets the constraints mentioned in section 6) can be reconstructed as a default graph in NDL, the inverse is not true. Most notably, default graphs that contain aggregation of reasons have no counterpart in NDL, for the reasons outlined in section 4. In other words, NDL is an extension of (a constrained version of) HDL. This means that D'Avila Garcez et al.'s approach (and Besold et al.'s deontic interpretation of this approach) are not suited to overcome the problems mentioned in section 4. Specifically, CILP-style neural networks do not allow for aggregation of reasons, since the logic programs to which they correspond do not allow for aggregation of reasons either. Moreover, since answer set programs (at least of the forms that D'Avila Garcez et al. and Besold et al. discuss) do not contain the syntax required to distinguish between (dis)favourers, attenuators/intensifiers, and disablers, CILP-style neural networks do not possess the structure to distinguish between these different types of reasons either. On the other hand, it must be noted that CILP-style neural networks have some advantages over the current version of NDL, specifically in that they allow for modelling cyclic answer sets (by means of connections from the output to the input layer). Besold et al. for instance exploit this feature to model contrary-to-duty obligations, which are not straightforwardly modelled in the current framework.

8.2 Inference Graphs

Pollock's ([28,29,30]) inference graphs consist of inference steps (which can be interpreted as propositions) that are linked together through edges. There are two types of edges, representing support and defeat relations, respectively. Originally, Pollock defined conclusions that can be drawn from inference graphs according to in/out-labellings, which are similar to extensions in HDL ([28,29]). However, in his [30], Pollock revised his framework to associate each proposition with a *justification strength* and each support edge with a *reason strength.* A conclusion's justification strength is a function of the strengths of its supporting propositions and the corresponding reason strengths, as well as the strengths of its defeaters. This feature is similar to NDL. Pollock's function, however, is very different from the functions that were presented in equations 1 and 2. First, while equation 1 means that multiple reasons for a proposition aggregate, Pollock rejects aggregation and identifies a conclusion's (*prima facie*) strength with the strength of its strongest argument only. This difference is a result of the fact that Pollock's interest is mainly in epistemic reasoning, rather than practical reasoning, and his believe that aggregation of reasons is more plausible in the domain of practical reasoning than in the domain of epistemic reasoning ([30], p. 246). While staying non-committal with regards to Pollock's claim about the non-aggregation of epistemic reasons, we agree that aggregation of different moral reasons is often an important step in ethical reasoning. Second, although Pollock importantly distinguishes between two types of defeaters, i.e. rebutting and undercutting defeaters, these defeaters affect a proposition's justification strength in the same way. In contrast, NDL makes a functional distinction between disfavourers (which are similar to Pollock's rebutting defeaters), attenuators and disablers (which are similar to Pollock's undercutting defeaters), which more naturally models commonsense moral reasoning.

9 Further Directions

Although NDL as defined in section 5 accounts for some patterns of reasoning that cannot be modelled by HDL, it is more limited in other respects, reflected in some of the limitations mentioned in section 6. The most pressing limitation of NDL is the restriction that default graphs are acyclic. Intuitively, defeat cycles are quite common in moral reasoning. For instance, if there are two available actions, A and B, only one of which can be performed, then it seems that the sentence 'A should be performed' attacks the sentence 'B should be performed', and *vice versa.* The most natural way to model this in NDL would be to add connections from n_A to n_B and *vice versa*, where both connections have a negative valuation (perhaps a valuation of -1). But this would introduce a cycle in the graph, which would mean that computing some nodes' valuation $V(n)$ according to equations 1 and 2 becomes less straightforward, since the valuations of n's input nodes may themselves depend on $V(n)$.

Future work may address this issue by introducing an iterative valuation function which, at each step τ_n, computes the valuations of all nodes simultaneously on the basis of the valuations at step τ_{n-1}, until it converges on a fixpoint. Such functions have been employed by D'Avila Garcez et al. ([17], ch. 3, [16], ch. 4), and also in some gradual approaches to Dung-style argumentation frames, which may contain cycles of argument defeat (e.g. [9,26,14]). The iterative function to determine the valuations of a cyclic default graph would be different from these earlier fixpoint functions, however, given the more complex ways in which different nodes can influence each others' activations in NDL.

Second, it would be useful to extend NDL to model more complex logical formulae and more complex defaults, compared to the simple formulae and defaults that conditions 1 and 3 in section 6 allow. For instance, it would be helpful to allow representation of negative premises and of other propositional operators, such as conjunction and disjunction. For this purpose, we may introduce 'operator nodes' that represent the valuations of logically complex sentences, and make the valuations of such nodes dependent on the valuations of their atomic constituent sentences. A natural approach, for instance, would be to have conjunction nodes take the minimum activation of all the nodes representing the conjuncts, and to have disjunction nodes take the maximum activation of the nodes representing the disjuncts.

Finally, future work may take advantage of the fact that default graphs in NDL have a strong structural similarity to artificial neural networks (ANNs). Like ANNs, default graphs consist of nodes and edges, such that the activation of each non-input node is a function of the activation of its input nodes, and the weights of the connecting edges. This similarity opens up the possibility to use learning algorithms that have been designed for ANNs in order to learn the weights of the connections of a default graph, similarly to D'Avila-Garcez et al. ([17,16]). For instance, we could provide a default graph with a set of training instances, representing descriptions of moral situations, and training labels, representing which act is morally right in each situation. Such labels may for instance be gathered by asking humans to judge the situations in question (see e.g. Guarini ([21]), who trained an ANN on the moral judgements of a group of students). Then, we could use backpropagation to iteratively update the default graph's weights on the basis of its error with regards to the training labels. It should be noted, however, that backpropagation is a gradient-based algorithm, i.e. it calculates the gradients of the error score with respect to each of a network's parameters through chained differential equations. This, however, requires each function within the network to be differentiable. However, the (inverse) step function used in section 6 is not differentiable at 0, and has a derivative of 0 for any other value. Future attempts to apply backpropagation to default graphs would need to consider this obstacle. Partially for this reason, the possibility to train default graphs like neural networks currently remains an uncertain aspiration.

If backpropagation-based learning would be possible for default graphs, this would allow such graphs to learn weights that represent the (average) moral intuitions of a group of people. This may be preferable to simply stipulating the connection weights, as we have done in earlier examples. At the same time, contrary to ANNs (such as the one trained by Guarini, [21]), default graphs would remain inherently interpretable, given that each node corresponds to a proposition or a default. This would be a substantial merit within the framework of machine ethics, where it is important that those affected by an automated decision can understand why the decision was made.

10 Discussion and Conclusion

The introduction of numeric weights, of course, invokes the question what these weights *mean*. Such weights cannot be interpreted in a probabilistic manner, as is the case for some other numeric frameworks in defeasible reasoning. The reason is that ought-statements, such as 'The car ought to steer to the side', are not descriptive propositions, and therefore it is unclear whether probabilities apply to them. Instead, the valuation of each output node in NDL seems to represent what can best be described as an act's *degree of rightness*. This is an idealisation of a commonsense notion that is sometimes expressed in sentences along the lines of 'I feel very strongly that I should...' or 'I feel somewhat obliged to...' (As a rough analogy, we may consider Bayesian probabilities as idealisations of sentences such as 'I know very certain that...')

At the same time, it must be recognised that the input nodes of default graphs do often represent descriptive propositions (such as 'Harm would occur if the car steers to the side'). The valuations of such nodes, then, can be (partially) dependent on the corresponding proposition's probability (they may also depend on other descriptive features, such as degree of harm). The way in which the 'descriptive valuation' of input nodes relates to the 'evaluative valuation' of output nodes is a question that cannot be decided here, since it requires extensive discussion of topics in moral philosophy. It should be noted, however, that there are reliable ways in which evaluative propositions depend on descriptive propositions: the higher the probability of harm, for instance, the more wrong an action is (*prima facie*). Any deontic default logic is based on these links, since each deontic default logic must allow for inferences from descriptive statements to evaluative statements. NDL makes these inferences more explicit by expressing them as numeric functions.

This paper has presented a numeric extension of HDL, where the strengths of both propositions and defaults are represented as numeric weights. It can be shown how NDL can reconstruct a (restricted but core) version of HDL, and also how it can be used to express patterns of moral reasoning that cannot be captured in HDL, most notably aggregation of reasons. For this reason, NDL offers interesting opportunities for combining the intuitions behind default logic (and nonmonotonic logic in general) with the affordances of the numeric modes of representation that are common in artificial systems. At the same time, it should be stressed that NDL is only a first step that can and should be extended in several ways to capture the full complexity of moral reasoning.

Acknowledgements

I wish to thank Jan Broersen for discussions of the ideas on which this paper is based, advice, and comments on earlier versions of the paper. I wish to thank Stipe Pandžic and Daniela Schuster for comments on an earlier version of the paper. I wish to thank Henry Prakken for literature recommendations. Finally, I wish to thank four anonymous reviewers for comments, suggestions, and literature recommendations.

References

[1] Allen, C., I. Smit and W. Wallach, *Artificial morality: Top-down, bottom-up, and hybrid approaches*, Ethics and Information Technology **7** (2005), pp. 149–155.

[2] Anderson, M. and S. L. Anderson, “Machine ethics,” Cambridge University Press, Cambridge, 2011.

[3] Anderson, M., S. L. Anderson and C. Armen, *Towards machine ethics*, in: *AAAI-04 workshop on agent organizations: theory and practice, San Jose, CA*, 2004, pp. 53–59.

[4] Arkoudas, K., S. Bringsjord and P. Bello, *Toward ethical robots via mechanized deontic logic*, in: *AAAI fall symposium on machine ethics*, The AAAI Press, Menlo Park, CA, USA, 2005, pp. 17–23.

[5] Besold, T. R., A. d. Garcez, K. Stenning, L. van der Torre and M. van Lambalgen, *Reasoning in non-probabilistic uncertainty: Logic programming and neural-symbolic computing as examples*, Minds and Machines **27** (2017), pp. 37–77.

[6] Bringsjord, S., K. Arkoudas and P. Bello, *Toward a general logicist methodology for engineering ethically correct robots*, IEEE Intelligent Systems **21** (2006), pp. 38–44.

[7] Cayrol, C. and M.-C. Lagasquie-Schiex, *Gradual valuation for bipolar argumentation frameworks*, in: *European conference on symbolic and quantitative approaches to reasoning and uncertainty*, Springer, 2005, pp. 366–377.

[8] Cervantes, J.-A., S. López, L.-F. Rodríguez, S. Cervantes, F. Cervantes and F. Ramos, *Artificial moral agents: A survey of the current status*, Science and Engineering Ethics **26** (2020), pp. 501–532.

[9] Da Costa Pereira, C., A. G. Tettamanzi and S. Villata, *Changing one’s mind: Erase or rewind? possibilistic belief revision with fuzzy argumentation based on trust*, in: *Twenty-second international joint conference on artificial intelligence*, 2011, pp. 164–171.

[10] Dancy, J., “Ethics without principles,” Oxford University Press, Oxford, 2004.

[11] Dong, H., B. Liao, R. Markovich and L. v. d. Torre, *From classical to non-monotonic deontic logic using aspic+*, in: *International workshop on logic, rationality and interaction*, Springer, 2019, pp. 71–85.

[12] Dung, P. M., *On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games*, Artificial Intelligence **77** (1995), pp. 321–357.

[13] Dunne, P. E., A. Hunter, P. McBurney, S. Parsons and M. Wooldridge, *Weighted argument systems: Basic definitions, algorithms, and complexity results*, Artificial Intelligence **175** (2011), pp. 457–486.

[14] Eğilmez, S., J. Martins and J. Leite, *Extending social abstract argumentation with votes on attacks*, in: *International workshop on theory and applications of formal argumentation*, Springer, 2013, pp. 16–31.

[15] Fuhrmann, A., *Deontic modals: Why abandon the default approach*, Erkenntnis **82** (2017), pp. 1351–1365.

[16] Garcez, A. S., L. C. Lamb and D. M. Gabbay, “Neural-symbolic cognitive reasoning,” Springer, Berlin Heidelberg, 2009.

[17] Garcez, A. S. d., K. Broda and D. M. Gabbay, “Neural-symbolic learning systems: foundations and applications,” Springer, London, 2002.

[18] Gill, M. B. and S. Nichols, *Sentimentalist pluralism: Moral psychology and philosophical ethics*, Philosophical Issues **18** (2008), pp. 143–163.

[19] Gordon, T. F., *Defining argument weighing functions.*, FLAP **5** (2018), pp. 747–773.

[20] Gordon, T. F. and D. Walton, *Formalizing balancing arguments.*, in: *COMMA*, 2016, pp. 327–338.

[21] Guarini, M., *Particularism and the classification and reclassification of moral cases*, IEEE Intelligent Systems **21** (2006), pp. 22–28.

[22] Horty, J., *Deontic modals: Why abandon the classical semantics?*, Pacific Philosophical Quarterly **95** (2014), pp. 424–460.

[23] Horty, J. F., “Reasons as defaults,” Oxford University Press, Oxford, 2012.

[24] Hunter, A., *A probabilistic approach to modelling uncertain logical arguments*, International Journal of Approximate Reasoning **54** (2013), pp. 47–81.

[25] Kagan, S., *The additive fallacy*, Ethics **99** (1988), pp. 5–31.

[26] Leite, J. a. and J. a. Martins, *Social abstract argumentation*, in: *Twenty-second international joint conference on artificial intelligence*, 2011, p. 2287–2292.

[27] Li, H., N. Oren and T. J. Norman, *Probabilistic argumentation frameworks*, in: *International workshop on theory and applications of formal argumentation*, Springer, 2011, pp. 1–16.

[28] Pollock, J. L., *Defeasible reasoning*, Cognitive Science **11** (1987), pp. 481–518.

[29] Pollock, J. L., *Justification and defeat*, Artificial Intelligence **67** (1994), pp. 377–407.

[30] Pollock, J. L., *Defeasible reasoning with variable degrees of justification*, Artificial Intelligence **133** (2001), pp. 233–282.

[31] Prakken, H., *Modelling accrual of arguments in aspic+*, in: *Proceedings of the seventeenth international conference on artificial intelligence and law*, 2019, pp. 103–112.

[32] Reiter, R., *A logic for default reasoning*, Artificial Intelligence **13** (1980), pp. 81–132.

[33] Ross, W. D., “The right and the good,” Oxford University Press, Oxford, 1930/2002.

[34] Thimm, M., *A probabilistic semantics for abstract argumentation*, in: *ECAI 2012*, IOS Press, 2012 pp. 750–755.

[35] Van Den Hoven, J. and G.-J. Lokhorst, *Deontic logic and computer-supported computer ethics*, Metaphilosophy **33** (2002), pp. 376–386.

[36] Wallach, W. and C. Allen, “Moral machines: Teaching robots right from wrong,” Oxford University Press, Oxford, 2008.

All-Things-Considered Ought via Reasons in Justification Logic

Federico L.G. Faroldi [1]

University of Pavia, Italy / CHAI, University of California at Berkeley
Pavia, Italy

Tudor Protopopescu

City University of New York
New York, USA

Abstract

We give a logical account of all-things-considered (*pro toto*) oughts via practical (*pro tanto*) reasons, adding to the deontic justification logic of [18] a very general relation of strength on sets of reasons. The agent can reason about reasons and then conclude what ought to be done, all-things-considered, given which reasons are stronger. In the first part of the paper we recall the deontic interpretation of justification logic. In the second part we show how to extend it to all-things-considered oughts. The resulting logic is explicit with regard to *pro tanto* reasons (which are expressed via terms), implicit with regard to the all-things-considered ought. In the final part of the paper we discuss some philosophical issues and ideas for future work.

Keywords: all-things-considered oughts, pro tanto reasons, justification logic, basic models, deontic logic, explicit modal logic

1 Introduction

Normative, or practical, reasons play a large, if not fundamental, role in contemporary normative theory. [2]

Normative reasons can be further distinguished into *pro tanto* and *pro toto* ones. [3] The former give some sort of initial justification for some obligation,

[1] federico.faroldi@unipv.it

[2] A reason can be thought of as a consideration in favor of or against something, at least for the scope of this paper. We remain non-committal about their metaphysical nature. Some take reasons to be primitive (or fundamental) for other normative notions (in the vicinity of this view are [39], [34], [13]). [36] takes reasons to be those things our rational capacities respond to; [27,29,28] take reasons to be evidence one ought to act in a certain way.

[3] A terminological point: in this paper we prefer *pro tanto* to *prima facie*, for the latter has unwanted epistemic connotations, but we use interchangeably *pro toto* and 'all-things-considered'.

whereas the latter justify an obligation once all relevant considerations for and against have been considered. *Pro tanto* and *pro toto* reasons are obviously related. The simplest and most naive idea is that one considers all the *pro tanto* reasons in favor of something, consider all the *pro tanto* reasons against something and, after some kind of aggregation, there is some process of weighing and the resulting reason (or aggregation of relevant reasons) is the *pro toto* reason for or against something. Given this very simple picture, one can easily see that *pro tanto* reasons will oftentimes conflict, whereas, if the aggregation and weighing processes are possible and consistent,[4] *pro toto* reasons won't – unless one's background theory is perhaps committed to inconsistent obligations for theoretical or pragmatic reasons.[5]

Related Literature Even without introducing more complicated metaethical ideas, reasons are seldom considered from a formal, and even less from a logical, point of view, as [26] confirms. There have been recent exceptions, like Horty's own work [25], which is mostly concerned about other normative notions like ought. [22] dedicates a whole section to what Goble calls *prima facie* obligations, which, under at least some readings, play a similar role to *pro tanto* reasons. Dietrich and List have been sketching a deontic logic [15] whose semantics is based on the reasons structures they define for their version of rational choice theory [14], reformulated with the normative interpretation given in [16]. As such, in their approach there is no way to deal with reasons in the object language and no logic of reasons themselves, nor a distinction between *pro toto* and *pro tanto* reasons. [18] is a recent attempt to explicitly deal with reasons, using the framework of justification logic (for a general introduction see [3,6]). Reasons are represented by justification terms, which can be manipulated and reasoned with, thus providing a first approximation to everyday practical reasoning.

Motivation There are two main advantages in using the justification logic framework to deal with normative reasons.

First, one can explicitly track which reasons are reasons for what, perform operations on them, and thus have a higher degree of accuracy in formal normative reasoning: every obligation has a source. Puzzles and paradoxes such as Ross' are very easy to identify and, under a plausible set-up, disappear.

Second, the framework allows for the hyperintensionality of obligation, namely that logically equivalent contents may not be normatively equivalent. In fact in general it is not the case that if $t{:}\phi$, "t is a reason for ϕ", and $\phi \equiv \psi$, then $t{:}\psi$, "t is a reason for ψ", see [18]. This also ensures a finer-grained formal approach to everyday normative reasoning that is currently unavailable in more standard approaches.

In [18], however, reasons were "flat", i.e. no ordering was imposed on them:

[4] For a skeptical argument on this front, see [17] and [12].

[5] Note that the logic we present here will allow for a conflict-tolerant all-things-considered ought, because we do not commit ourselves to the existence of a consistent aggregation and weighing process. For more on this, see Sect 6.

this means that no reason (or set of reasons) is more important than another, and therefore one cannot conclude what ought to be done in case of conflicts. Without doubt, there are several ways to pass from *prima facie* (or *pro tanto*) reasons, oughts or obligations to the all-things-considered (or *pro toto*) version. One traditional idea, for instance, is that the final, actual obligation is just an (or the) undefeated *prima facie* one(s). [6] Another idea is that one ought to do whatever one has most reason to do. [7]

In this paper we propose to make the framework developed in [18] a bit more adequate to real-world agents, by adding a way of comparing (simple and complex) reasons, thus making it possible to have all-things-considered obligations. [8]

The plan of the paper is as follows. In Sect. 2 we explain the philosophical ideas on which our formal account is based. In Sect. 3 we briefly introduce justification logics and discuss their intended deontic interpretation, focusing on a specific system in Sect. 4. In Sect. 5 we show how to extend it to all-things-considered oughts. The resulting logic is explicit with regard to *pro tanto* reasons, but implicit with regard to the all-things-considered ought. In the final section of the paper, Sect. 6, we put forward some philosophical remarks and point out some ideas for future work.

2 Pro tanto reasons and all-things-considered ought

This section will informally introduce and discuss the ideas used later in the paper.

Example 2.1 Suppose you just promised a friend to meet at 3pm in the main square. Your promise is a reason for why you ought to be in the main square at 3. But further suppose that, on your way to the meeting, a crazy biker, after a hazardous maneuver, falls and seems quite seriously wounded, and you are the only passerby. This seems a reason why you ought to call an ambulance and wait for the biker to be succoured. Now, let's agree that, if you stay and help the biker, you won't make it to your meeting: one excludes the other. What to do?

In this very simple case it is quite natural to say that you ought to stay and help the biker. [9] But why? Let's try to generalize: what seems to be

[6] [10, p. 149], [11, p. 125]: an agent ought to perform an action just in case there is an undefeated reason for the agent to perform that action.

[7] In favor of the thesis that an agent ought to perform an action just in case the reasons that favor that action outweigh the reasons disfavoring it cf. [40, p. 130] For a comprehensive and critical survey, see [22].

[8] Once one reads reasons as *prima facie* or *pro tanto* obligations or duties, the underlying philosophical idea is of course not new, and in modern times can be retraced at least back to [38]. Formal investigations started at least with [1]. For other more recent formal treatments of this issue, see at least [9], [23], [24], and, more recently and within the adaptive logic framework, [41]. For deontic logics based on reasons (based on preferences), without taking into account the strength of obligations, see also [33,32].

[9] Whether this enjoins further obligations on your part (such as letting your friend know) is

going on here is not that, given the factual details of the situation, you just considered your two obligations in isolation, as it were; rather there seems to be some sort of reasoning performed thanks to and on the reasons for the two obligations. Somewhat tentatively we might want to say that the reason for staying and helping is stronger, or more important, than the reason for going to the meeting. This is easily shown if we manipulate our example a little bit by adding other reasons for the obligation to go to the meeting, which are stronger than that for staying and helping the biker, or whose combination or aggregation with the reasons already existing makes a stronger case; for instance, if the failure to meet with your friend results in the injury of more people.

How to account for that? There are two main views: the weight-first view, and the ordering-first view. On the former view, the weight of individual reasons is given, and the ordering is then derivative on the ordering on weights (possibly the natural ordering on real numbers).[10] On the latter view, the first natural idea that springs to mind is to impose some sort of ordering on the reasons, something akin to a preference ordering on options in decision theory, but whose intended interpretation is normative, rather than motivational or descriptive, and, in case some conditions are respected, some weights or other can then be assigned. However, there are many well-known problems with these ideas, having to do with either the requirement that this ordering be complete, or with transitivity, or with some sort of separability condition, not to mention the strong conditions required to prove a representation theorem into quantitative structures.

What we propose is to generalize this idea and to use a generic binary relation on sets of reasons for actions, rather than on individual reasons.[11] This relation is not required to be an ordering.[12]

Such a relation seems to capture what normally happens in everyday reasoning, where you consider all the reasons in favor of staying and helping the biker, and all the reasons not to (including the reasons favoring going to the meeting), and then we conclude what we ought to do, all things considered, given which of the two collections is more important.

Defining this relation on sets of reasons, rather than on individual reasons, has an advantage: such an account lets you be non-committal about how reasons aggregate, and therefore the above-mentioned problems can be put aside. In particular we don't force an atomistic view of reasons (according to which they have the same weight no matter the context, e.g. which other reasons are present or absent) nor a holistic view of reasons. If you want to track

a rather interesting question we don't tackle here.

[10] For several objections to this view, see [17].

[11] For a similar idea, although in a slightly different context, see [40, Ch. 7]; and [16] for a decision-theoretic context. Using a relation on sets of reasons, rather than a binary relation on individual reasons was also criticized in [17].

[12] [25] defines a partial ordering on defaults of the form $X \to Y$, where X could perhaps serve as a reason in our sense, given it meets further conditions.

the reasoning of agents and, ultimately, give a logical account, certain more metaphysical questions can be productively abstracted away.

These ideas are realized by adding to the semantics a binary relation $>$ on sets of reasons, which are represented by terms in the logic. Moreover, since the semantical interpretation of terms we adopt for the purposes of this paper is that a term is a set of formulas (namely, the set of formulas it justifies), it is easy to see that this binary relation is a relation between sets of sets of formulas: $>\subseteq 2^{Tm} \times 2^{Tm}$.

Is this relation an ordering? No: the rationale is to be as general as possible, for two reasons: first, to avoid certain technical problems of standard orderings; second, to have a certain amount of freedom to be able to plug in, in this logic, your preferred normative theory, or an approximation thereof. For instance, it is quite sensible to think that, were we to require the binary relation to be a preference ordering (i.e. transitive and complete), we would be then reasoning with a broadly teleological (if not consequentialist) normative theory. One can also impose a separability condition on the relation, in order to recover additive theories of value.[13] Not imposing further conditions on the binary relation on sets of reasons fulfills another desideratum: generality.[14] In fact it is compatible with the thought that, in different worlds, the characteristics of the "ordering" are different, perhaps because the background normative theory is different.

Given this binary relation on sets of reasons, an obligation to ϕ is now all-things-considered, or *pro toto*, which we will write standardly as $\mathcal{O}\phi$ just in case that the set of reasons for ϕ is more important or stronger than the set of reasons for $\neg\phi$, which provides the intuitive reading of formulas like $(t{:}\phi \succ s{:}\neg\phi)$.[15]

This understanding presupposes that the situations represented by ϕ and by $\neg\phi$ are exclusive. This, in a sense, respects our underlying intuitions: *pro tanto* reasons can conflict; in case of conflict, however, there will be a *pro toto* obligation only if you can find, among the conflicting reasons, a set of them that clearly supports an outcome and is more important than the others. If this cannot be found, then there is no all-things-considered obligation.

[13] For several options in this respect, see [16] and later on in this paper.

[14] But won't technical problems of standard orderings come back via the plugged-in normative theory, if the chosen normative theory employs one of those standard orderings? Yes. But this highlights that the problems are to be imputed to the background normative theory, rather than to the formalization. In this sense we can see how generality, whose other side of the coin is underspecification, can be of theoretical help.

[15] The fact that the set of reasons for ϕ is more important than the set of reasons for $\neg\phi$ amounts to an all-things-considered obligation to ϕ gives an indication of what this "arbitrary" relation captures, also in the general case. While such an approach raises modeling questions, the problem is that the informal weighing relation between reasons (as discussed in the informal ethics literature) is not well understood in the first place (for some evidence for this claim, see [17, Ch. 6]).

3 Logics for *pro tanto* reasons

The framework of Justification Logic [3,6], a.k.a. explicit modal logic, offers a formal setting in which we can represent reasons explicitly allowing us to represent reasoning about the reasons for which something is obligatory. We offered such an account in [18]. In this section we briefly summarize the basics.

3.1 Explicit Deontic Logic

The generic deontic reading of a modal formula $\Box A$, or $\mathcal{O}A$, is

A is obligatory.

Presumably in each case there is some reason for this obligation, but the language of standard deontic logic is not capable of denoting it explicitly.

In the explicit modal language of Justification Logic modalities, $\Box$'s, are decomposed into terms, t, denoting the specific reason why a proposition is justified, proved, known, believed etc. In a deontic setting this means that formulas of the type $\mathcal{O}A$ are replaced with formulas of the type

$$t{:}A.$$

Some possible readings of this are

t is a reason why A is obligatory

or

t is a reason to do A

or

t is a reason for A being the case

or

you ought to do A because of t. [16]

Readings of conjunctions and disjunctions are obvious: $t{:}(A \wedge B)$ may be read as *t is a reason why 'A and B' are obligatory* or as *t is a reason to do both A and B*. Likewise $t{:}(A \vee B)$ might be read as *you ought to do either A or B for reason t*. The reading of the material conditional is standard, although at this point one can ask whether it would not be better to have genuine conditional obligations, perhaps represented as $A \rightarrow t{:}B$. These points are of course going to be relevant when trying to formalize so-called contrary-to-duty oughts, but we don't think having reason terms substantially alters these problems, which are directly inherited from, implicit, standard deontic logic. [17]

[16] At this point, as it is clear, we are not committed to an ought-to-do or an ought-to-be reading. We do not exclude, however, that an enrichment of the framework allows for such a distinction. We also do not claim that the above possible readings are necessarily mutually consistent; different readings will, presumably, give rise to different logics.

[17] [37] tackles conditional obligations in the framework of justification logic.

3.2 Operations on Terms

The other key feature of an explicit modal language is the possibility of defining operations on terms. In general operations on terms represent some kind of reasoning with our reasons, in order to obtain new reasons. The most fundamental operation is *application*, which can be thought of as representing a step of *modus ponens*.

Assume that s is a reason for A, $s{:}A$, and t is a reason for $A \to B$, $t{:}(A \to B)$, for the sake of concreteness say that s and t are proofs in a formal system with the last line A and $A \to B$ respectively. If one applies *modus ponens* to t and s one thereby obtains a proof of B; this proof is denoted by $(t{\cdot}s){:}B$.

Application is the main operation for reasoning with reasons, not just formal proofs. These observations, hence, justify and are encoded in the *application principle*

$$t{:}(A \to B) \to (s{:}A \to (t{\cdot}s){:}B)$$

Application is the only operation we are concerned with here, but it is not the only operation possible. For example the first justification logic, the Logic of Proofs [2], the explicit counterpart of $\mathsf{S4}$, the system $\mathsf{JT4}$ of [3,6], contains two operations in addition to application: *plus*, + – which takes two proofs, t and s and returns a proof $(t+s)$ which proves anything that either t or s proves – and *proof checker*, !, which takes a proof t and returns a proof $!t$ which is a proof that t is indeed a proof. [21] and [7] contain several examples of other possible operations and discuss the issue in a general setting. The basic explicit deontic logic proposed in [18], the system $\mathsf{NRL}_{\mathcal{CS}}$, contains plus in addition to application; it also considers that ! and a *reflection* operation, ‡ – which takes a deontic reason t and returns a higher order reason $‡t$ to satisfy the obligation that t enjoins – might make sense in a deontic context.

In what follows we take only the minimal system of justification logic, J^-, in order to have in place most of the technical machinery needed to properly set up our discussion of *pro tanto* reasons and all-things-considered obligations later on.

4 The Minimal Justification System J^-

Definition 4.1 [Basic Explicit Language]

The basic explicit language contains the following items:

1. The language of classical propositional logic;
2. *Variables*: $x, y, x_1, x_2 \ldots$;
3. *Constants*: $a, b, c, c_1, c_2 \ldots$;
4. *Application*, $\cdot$.

Formulas are built up according to the following rules:

- *Terms*: any variable or constant is a term; if t and s are terms so is $(t{\cdot}s)$.
- *Formulas*: As for classical logic, and additionally if A is a formula and t a term then $t{:}A$ is a formula.

Now we define the minimal system of justification logic J_0^-.

Definition 4.2 [J_0^-] J_0^- consists of the following axioms and rules:

J0. Axioms of classical propositional logic;

J1. $t{:}(A \to B) \to (s{:}A \to (t{\cdot}s){:}B)$.

- Modus Ponens

4.1 Variables and Constants

A key feature of justification logics is that terms can be of two types: *variables* and *constants*. Variables represent arbitrary reasons, while constants represent reasons for assumptions, and in particular for axioms. Constants represent specific "atomic" reasons, i.e. reasons which are simply accepted as justifying what they do.

Deontically, reasons for logical axioms can be understood to represent how logic-respecting one's obligations are. For instance for a constant c, $c{:}((A \wedge B) \to A)$ might be read as *c is a reason to do A on condition that A ∧ B is done.* Another example is $c{:}(A \to (A \vee B))$, which can be read as *c is a reason to do A∨B on condition that A is done.* The first might seem plausible, because rather trivial, while the second might be more implausibile.

Fortunately, the framework of Justification Logic is very flexible when it comes to the choice of basic assumptions and their justification, as represented by constants. Technically this is known as a *constant specification* (see definition 4.3), the set of assumptions justified by constants. To anticipate, a constant specification may range from being empty (no assumption is justified) to being total (any constant is a reason for any assumption).

Several deontic puzzles depend on the idea that axioms or tautologies are obligatory, for instance by derivation with the necessitation rule. The mechanism of a constant specification allows a great degree of control over this kind of reasoning, and is an important ingredient in the analysis of some standard deontic puzzles. Given that many puzzles in standard deontic logic center on, or involve, necessitation or the assumption that some logical principle is obligatory, it is conceivable that the most realistic kinds of explicit deontic systems will have empty, or very limited, constant specifications.

In J_0^- no axiom is justified, i.e. no formulas of the form $t{:}A$ are derivable where A is an axiom of J_0^-. If one wishes to be able to derive conclusions of the form $t{:}A$ one needs to assume some axioms are justified. This is one of the purposes of a constant specification.

Definition 4.3 [Constant Specification]

A *constant specification*, $\mathcal{CS}$, is a set of formulas of the form $\{c_n{:}c_{n-1}{:}\ldots c_1{:}A\}$ where $c_n, c_{n-1}, \ldots, c_1$ are constants and A is an axiom. It is assumed that if $c_n{:}c_{n-1}{:}\ldots c_1{:}A \in \mathcal{CS}$ then so is $c_{n-1}{:}\ldots c_1{:}A$.

A constant specification $\mathcal{CS}$ keeps track of the assumptions, i.e. the axioms of the given system, which are considered justified. The members of a given $\mathcal{CS}$ can be added to J_0^- as additional axioms to obtain the system $\mathsf{J}_{\mathcal{CS}}^-$. Note that J_0^- is $\mathsf{J}_{\mathcal{CS}}^-$ with $\mathcal{CS} = \emptyset$.

In general, what is in a given $\mathcal{CS}$ is up to the user. A $\mathcal{CS}$ may be a finite set, stating that only some of the axioms of the system are justified. An important class of $\mathcal{CS}$'s are the *axiomatically appropriate* ones. A $\mathcal{CS}$ is axiomatically appropriate when each axiom of the system, including those in the $\mathcal{CS}$ itself, is justified, i.e. if $c_n{:}c_{n-1}{:}\ldots c_1{:}A \in \mathcal{CS}$ then $c_{n+1}{:}c_n{:}c_{n-1}{:}\ldots c_1{:}A \in \mathcal{CS}$. A $\mathcal{CS}$ is *total* when for every axiom and any constant $c_1, c_2, \ldots c_n,\ \ c_n{:}c_{n-1}{:}\ldots c_1{:}A \in \mathcal{CS}$.

A $\mathcal{CS}$ makes possible certain interactions between terms and the logical connectives.

Definition 4.4 [J^-]

The system J^- is the system J_0^- with the additional rule of Axiom Necessitation:

$$\frac{\vdash A}{\vdash c_n{:}c_{n-1}{:}\ldots c_1{:}A}$$

where A is J0 or J1 and $c_1, c_2, \ldots c_n$ are any constants.

J_0^- coincides with $\mathsf{J}^-_{\mathcal{CS}}$ with a total $\mathcal{CS}$.

A *ground* term is a term built up entirely from constants. Given an axiomatically appropriate $\mathcal{CS}$ or axiom necessitation the rule of *constructive necessitation* is derivable: $\frac{\vdash F}{\vdash t{:}F}$ for a ground term t. This rule follows from a more general property of systems with an axiomatically appropriate $\mathcal{CS}$: they are able to internalise their own proofs: If $A_1 \ldots A_n, y_1{:}B_1 \ldots y_n{:}B_n\ \vdash F$ then for some term $p(x_1 \ldots x_n, y_1 \ldots y_n)$

$$x_1{:}A_1 \ldots x_n{:}A_n, y_1{:}B_1 \ldots y_n{:}B_n\ \vdash p(x_1 \ldots x_n, y_1 \ldots y_n){:}F$$

Since a number of significant puzzles in standard deontic logic involve necessitation, or the obligatoriness of logical principles, we mention these options to give a sense of the degree of control possible over such assumptions. [18]

5 Logics for *pro toto* reasons

We now come to the novel developments. There are two additions to the language of basic justification logic we introduced in the previous section. First, the importance or strength relation $\succ$; second, the implicit modality $\mathcal{O}$ (i.e. $\mathcal{O}$ is unrealized by a set of terms, see 3.1), expressing all-things-considered obligation. All-things-considered obligation depends on the balance of reasons, namely, if we have an all-things-considered obligation then we know that the reasons in favor are more important, or stronger, than the reasons against.

5.1 Syntax

To get the logic PTJ^-, the logic for *pro toto* reasons, we modify the system J^- above by expanding the language, definition 4.1, with the two following operations:

Definition 5.1 [Language of PTJ^-]

[18] For more details and proofs see, for instance, [2,3,6,19,21].

5. *"Stronger Than"*: $\succ$;
6. *Obligation*: $\mathcal{O}$.

Accordingly, we expand the definition of formulas thus:

- *Formulas*: if A is a 'purely propositional' formula, i.e. not containing any occurences of $\mathcal{O}$, $\succ$ or any term t, then $\mathcal{O}A$ is a formula.
- If A and B are formulas not containing any occurence of $\mathcal{O}$, $\succ$ or any term t, then $t{:}A \succ s{:}B$ is again a formula.

Note that, $t{:}A \succ s{:}B$ and $\mathcal{O}A$ behave like atomic formulas.

Definition 5.2 [PTJ^-] PTJ^- consists of the following axioms and rules:

J0. Axioms of classical propositional logic;

J1. $t{:}(A \rightarrow B) \rightarrow (s{:}A \rightarrow (t{\cdot}s){:}B)$;

J2. $\mathcal{O}A \rightarrow (t{:}A \succ s{:}\neg A)$.

- Modus Ponens
- Axiom Necessitation

5.2 Basic Models

We will now introduce a semantics known as basic models (cf. /e.g. [4]). The principle advantage of basic models is that they give a precise answer to the question 'what is a reason?': a reason is represented by a set of formulas, namely, those it supports. Possible-world models (i.e. Fitting semantics) instead treat justification terms as undefined primitive objects. Another advantage is that basic models keep separate the question of the truth of A from that of $t{:}A$: whether it is true that t is a reason for A is a separate question from whether it is true that A. Possible-world models do not separate these questions completely; the truth of $t{:}A$ depends on the truth of A at some (other) states. Basic models treat A and $t{:}A$ as distinct formulas; indeed they treat formulas of the form $t{:}A$ as atomic.

Basic models are built from two sets, that of *justification terms*, Tm, and that of formulas Fm (built in the usual way from propositional atoms using Boolean connectives and also via justification terms of the form $t{:}A$). In a basic model formulas are interpreted as truth values, and terms are interpreted as *sets of formulas*; i.e. a term is just the set of formulas for which it serves as a justification.

In order to interpret the application operation we first need to define the set of formulas $\Phi \rhd \Gamma$ which is the result of applying *modus ponens* to the members of sets of formulas Φ and Γ.

Definition 5.3 [$\rhd$] For sets of formulas Φ and Γ and formulas A and B,

$$\Phi \rhd \Gamma =_{df} \{B \mid A \rightarrow B \in \Phi \text{ and } A \in \Gamma\}.$$

With this we now define a basic model for the logic of all-things-considered obligations:

Definition 5.4 [PTJ$^-$ Basic Model]
A PTJ$^-$ basic model $(*, >)$ consists of the following:

1. An interpretation of the elements of the set of atomic formuals, At, to truth values, $\{0, 1\}$, and the elements of Tm to sets of formulas, i.e.
$$* : At \mapsto \{0, 1\}$$
and
$$* : Tm \mapsto 2^{Fm}.$$
2. $s^* \rhd t^* \subseteq (s{\cdot}t)^*$ [19]
3. $A \in t^*$ for any conclusion $t{:}A$ of axiom necessitation.
4. A binary relation $>$ between sets of sets of formulas: $> \subseteq 2^{2^{Fm}} \times 2^{2^{Fm}}$.

Definition 5.5 [Truth in a PTJ$^-$ basic model]
Truth in a PTJ$^-$ (indeed any) basic model is defined inductively:

1. $(*, >) \Vdash P$ iff $P^* = 1$, for atomic P.
2. Boolean conditions for the propositional connectives.
3. $(*, >) \Vdash t{:}A$ iff $A \in t^*$, for any $t \in Tm$ and any $A \in Fm$.
4. $(*, >) \Vdash t{:}A \succ s{:}B$ iff $\{t^* | A \in t^*\} > \{s^* | B \in s^*\}$, for all $t, s \in Tm$.
5. $(*, >) \Vdash \mathcal{O}A$ iff $\{t^* | A \in t^*\} > \{s^* | \neg A \in s^*\}$, for all $t, s \in Tm$ and $A \in Fm$.

Definition 5.6 [Consequence] $\Gamma \Vdash A$ iff for every $(*, >)$ and for every $B \in \Gamma$, if $(*, >) \Vdash B$ then $(*, >) \Vdash A$.

Theorem 5.7 (Soundness) *If* $\mathsf{PTJ}^- \vdash F$ *then* $(*, >) \Vdash F$ *for any* PTJ^- *basic model.*

Proof. By induction on derivations.
The Boolean cases are standard. Let us check the justifcation axioms:

1. F is $t{:}(A \to B) \to (s{:}A \to (t{\cdot}s){:}B)$. Assume $(*, >) \Vdash t{:}(A \to B)$ and $(*, >) \Vdash s{:}A$, hence $A \to B \in t^*$ and $A \in s^*$. Hence $B \in t^* \rhd s^*$, and $B \in (t{\cdot}s)^*$; and so $(*, >) \Vdash (t{\cdot}s){:}B$.
2. F is $\mathcal{O}A \to (t{:}A \succ s{:}\neg A)$. Assume $(*, >) \Vdash \mathcal{O}A$, hence for any $t, s \in Tm$ and any $A \in Fm$ $\{t^* | A \in t^*\} > \{s^* | \neg A \in s^*\}$, and hence $(*, >) \Vdash t{:}A \succ s{:}\neg A$.

□

Theorem 5.8 (Completeness) *If* $(*, >) \Vdash F$ *for any* PTJ^- *basic model then* $\mathsf{PTJ}^- \Vdash F$.

Proof. By constructing a canonical model $(*^c, >^c)$ as in Theorem 3.8 in [30], with the following additions for the ordering clause. For a maximally PTJ^--consistent set of formulas Φ

1. $t{:}A \succ s{:}B \in \Phi$ iff for all t and s, $\{t^* | A \in t^*\} >^c \{s^* | B \in s^*\}$.

[19] In case we have $+$: $s^* \cup t^* \subseteq (s + t)^*$.

2. $\mathcal{O}A \in \Phi$ iff for all t and s $\{t^*|A \in t^*\} >^c \{s^*|\neg A \in s^*\}$.

By definition $>^c$ is a binary relation between sets of formulas, hence item 4. of definition 5.4 is satisfied, and the canonical model is a PTJ^- model.

As usual the Truth Lemma, for any formula F, $F \in \Phi \Leftrightarrow (*^c, >^c) \Vdash F$, is proved by induction on the complexity of formulas. We add the following two cases:

F is $t{:}A \succ s{:}B$.

$\Rightarrow$: Assume $t{:}A \succ s{:}B \in \Phi$, then $\{t^*|A \in t^*\} >^c \{s^*|B \in s^*\}$, and hence $(*^c, >^c) \Vdash t{:}A \succ s{:}B$.

$\Leftarrow$: Assume $t{:}A \succ s{:}B \notin \Phi$, then for some t^* and s^* $\{t^*|A \in t^*\} \not>^c \{s^*|B \in s^*\}$, hence $(*^c, >^c) \nVdash t{:}A \succ s{:}B$.

F is $\mathcal{O}A$.

$\Rightarrow$: Assume $\mathcal{O}A \in \Phi$, then $\{t^*|A \in t^*\} >^c \{s^*|\neg A \in s^*\}$, and hence $(*^c, >^c) \Vdash \mathcal{O}A$.

$\Leftarrow$: Assume $\mathcal{O}A \notin \Phi$, then for some t^* and s^* $\{t^*|A \in t^*\} \not>^c \{s^*|\neg A \in s^*\}$, hence $(*^c, >^c) \nVdash \mathcal{O}A$.

□

Example 1, continued The following is a basic model representing our biker example from above:

i) A = you meet your friend at 3pm

ii) B = you help the biker

iii) t = you made a promise to your friend

iv) s = you are the only passerby and the biker is seriously wounded

v) A and B are mutually exclusive, in particular $B \to \neg A$, s is a reason for $\neg A$

vi) $(t{:}A)^* = 1$, $(s{:}B)^* = 1$ $(s{:}\neg A)^* = 1$, $t^* = \{\{A\}\}$, $s^* = \{\{\neg A, B\}\}$, $s^* > t^*$.

We can now conclude that $\mathcal{O}\neg A$.

In the last section we are going to expand these considerations by taking into account some connections to recent choice-theoretic developments.

6 Philosophical Remarks and Future Work

Conflicting Oughts The current approach is extremely flexible, maybe even too flexible. If no further conditions are imposed on $>$, the relation between sets of reasons, it may very well happen that (i) there is no all-things-considered ought: the relation in fact is not required to be connected; (ii) there might be all-things-considered oughts for inconsistent things.[20] This fact raises two questions. In one hypothesis, we could change the understanding of *pro toto* obligations, by postulating that a *pro toto* obligation is no more than an un-

[20] In particular, unless $>$ is assumed to be asymmetric, one can build a model where $\mathcal{O}A \wedge \mathcal{O}\neg A$ is true. This, however, does not result in $\mathcal{O}(A \wedge \neg A)$ as in standard deontic logic.

defeated *pro tanto* reason, where the defeasibility, as it were, in our present approach is simply captured by checking that there are no reasons that are more important or stronger. Therefore, under such an understanding, the fact that two particular sets of reasons are unrelated does not matter for *pro toto* obligations, because we have a way of generating them automatically from the *pro tanto* reasons we have. Such position is neither new nor satisfactory, and will just set us back at the beginning of our inquiry. In another hypothesis, such incompleteness should be embraced insofar as it is an accurate description of our messy normative life. Suppose we don't have *pro toto* obligations, because all *pro tanto* reasons are indecisive. Then other practical, extra-normative methods will be needed to take a decision about what to do. Consider, for instance, a relation that happens to be cyclic: $a{:}(p \wedge \neg q), b{:}(q \wedge \neg r), c{:}(r \wedge \neg p)$ and $a^* > b^*, b^* > c^*, c^* > a^*$. In this case, there is simply no all-things-considered obligation. If we think that such a situation is unacceptable, then we could require that $>$ be connected.

Reason aggregation There is an obvious question to answer, that is, how the ordering on sets of reasons works and should work in the non-atomic case. The ordering would be separable if given c disjoint from a and from b, if $a > b$ then $a \cup c > b \cup c$, where the union stands for the relevant notion of sum, etc. It is well-known that additivity is a special case of separability, so all additive representations are separable.

In our practical reason case, it is easy to see that, if we had $+$, given $s^* \cup t^* \subseteq (s+t)^*$, our ordering would indeed be separable. Without $+$, however, assuming that, in the metalanguage, you have complex reasons (sets of formulas) that are the union of others, their place in the ordering will have to be specified case by case.

This, moreover, will be relevant to all-things-considered obligations just in case all the reasons we aggregate are indeed for or against the same proposition. We cannot, however, just assume something like $t{:}A \wedge s{:}A \rightarrow (t+s){:}A$, for it's not clear that two reasons for the same thing when taken separately, are still going to be a reason for that thing when taken together.

Pro tanto *vs* Partial Reason What about partial reasons?

For the time being, let's put aside whether partial reasons are indeed reasons. We are going to talk about partial considerations in favor or against something. Roughly, a partial consideration in favor of something is a consideration that, taken alone, is not sufficient to establish an obligation. However, it may be the case that several partial considerations taken together could support a full-fledged obligation. As a very simple example, the fact that almost everyone is going to be there might be a partial consideration for me to go to the department party. Taken by itself, that's not enough: were for that fact only, I won't be bothered to think I have a full-fledged reason to go the party. Likewise, I know a dear friend is going to be there. Taken by itself, that's not enough, were for that fact only, I won't be bothered to think that I have a full-fledged reason to go to the party. They're both partial considerations in favor of going. But perhaps taken together they are a reason (perhaps *pro* *tanto*, if not *pro toto*) to go. So far, there is no way to express partial considerations in our system. As a matter of fact, *pro tanto* reasons are full reasons. Here is a possible way to distinguish between *pro tanto* reasons and partial considerations (reasons). Suppose that there is only one *pro tanto* reason, r. Then, r would become an all-things-considered reason, a *pro toto* reason, an obligation to ϕ. However, suppose that there is only one partial consideration for ϕ. Then, in the absence of other considerations or reasons for or against ϕ, there won't be a reason or an obligation to ϕ.

Here's how we can handle partial considerations. Recall that, in the semantics, the interpretation of a term is a set of formulas, i.e. the set of formulas it justifies. Under normal conditions, operations on terms are specified recursively. Let's take a concrete example, i.e. the plus operation.

The interpretation of $(t + s)$, i.e. $(t + s)^*$, is just the union of the interpretations of the individual terms, i.e. $t^* \cup s^*$. Now it's clear, for elementary set-theoretic facts, that a given formula A cannot be in $t^* \cup s^*$ without being either in t^* or in s^*. But that's exactly what we need to express partial considerations. Let's consider again the example above. Set $a :=$ 'almost everyone is going to be at the party'; $b :=$ 'my friend is going to be there' and $C :=$ 'I go to the party'. Now, neither of a or b, taken alone, is a reason to go to the party, C. However, let's suppose that together they are.

Thus, it seems that $a + b$ is indeed a reason for C, without a or b being reasons as well. So it can't be the case that $(a + b)^* = \{C\}$ without $a^* \cup b^* = \{C\}$.

The obvious workaround is to define a new operation analogous to +, $\pm$, directly specifying an interpretation of "aggregated" reasons.

More formally, $(t \pm s)^* \neq t^* \cup s^*$, but is understood primitively. In this sense, we modify the clause for terms in this way: $\tau^* = 2^{Fm}$, or $^* : Tm \to 2^{Fm}$, where $\tau \in Tm$ is of the form t, $t \cdot s$ or $t \pm s$.

Such a set-up is different than defining the plus operation as a partial operation, because given that $t \pm s$ has an interpretation, it is not required that either t or s, taken alone, do.

Moreover, such a clause is perfectly compatible with another feature of our system. In particular we defined an ordering relation on sets of terms, and we asked whether this ordering is separable. If $\pm$ were defined point-wise, like + is, then this seems to force the ordering to be separable (because $\pm$ would behave additively), whereas by defining $\pm$ primitively, the position of a given $t \pm s$ in the ordering would not depend on the positions in the ordering of the individual factors.

Defeaters In line with our first hypothesis in the preceding paragraph, we may ask if we can provide a more precise understanding of defeaters. One of the most prominent formal theories of oughts taking into account (what he calls) reasons is, as already mentioned, Horty's (see in particular [25]). Horty's work is based on default theory. He also has an ordering on defaults, representing their priority. As it is well known from this tradition, there are two kinds of defeaters: *rebutting* defeaters, i.e. additional stronger reasons for a conflicting conclusion, and *undercutting* defeaters, ones that somehow impact the capacity of a reason to be a reason at all. In the current framework, a rebutting defeater A for $t{:}\phi$ can be of the form, given $t{:}\phi$, $A \to s{:}\neg\phi$, with $\{s\} > \{t\}$. An *undercutting* defeater B, given $t{:}\phi$, can be of the form: $B \to \neg t{:}\phi$. It is in this latter sense that Raz's exclusionary reasons can be understood (e.g. [35]).

The importance of reasons and reasons for importance Open remains the question of what to do with reasons *for* priority claims, i.e. possible reasons of the form $t{:}(s{:}A \succ u : B)$, which, in a radically different framework, are allowed by [25, Ch. 5] and are excluded in the present framework. Such a feature would allow *practical* reasoning about priorities. However, while it seems to be conceptually hard to make sense of iterated or "second-order" reasons for reasons, [21] it may be appropriate to interpret reasons for priority claims in deontic terms as intensifiers and attenuators, to the extent that intensifiers and attenuators are themselves reasons (and not extranormative considerations, as e.g. [13] and [8] think).

Wrong reasons Since there is now a way to express *pro tanto* reasons, which, at least in the presence of other reasons, do not generate all-things-considered obligations, it might make sense to ask whether there is a way to express that something is a wrong reason. We do not engage with the philosophical debate. The most natural way to account for this thought is the following. Assuming that right and wrong reasons are full-fledged reasons, and that they are disjoint, one can simply introduce two disjoint sets of terms: one for the right reasons, and one for the wrong reasons, see [5] for an approach in the context of epistemic logic. Moreover, one restricts the ordering on sets of terms to subsets of the right reasons, because wrong reasons do not contribute to form all-things-considered obligations. More controversially, one has also to modify the clause for obligations, restricting the *pro tanto* reasons contributing to the all-things-considered obligation to the ones which are in the set of the right reasons. As a matter of fact, one could think that such a modification is useless, for the ordering is already defined only on the right reasons. However, assuming that there is only one (right) *pro tanto* reason, it will be vacuously "better" than all others, thus becoming our *pro toto* obligation. This unfortunately would also happen if there is only one *wrong pro tanto* reasons, which might then become our *pro toto* obligation, rather than there being no obligation whatsoever, as intuition demands.

Subset semantics and quantification over terms We conclude with two technical ideas for future work.

The first has to do with the semantics, and involves developing a semantics for the system discussed in this paper not based on basic models, which still retains a syntactic feeling, but rather on the subset semantics developed by Studer et al. [37], where terms are assigned not just a set of formulas, but a set of worlds, and thus have themselves a semantical content.

[21] For a general discussion about iterated deontic reasons, see [18].

The second idea has to do with quantifying over terms in the manner of [20]. In the present context, such an extended system would be able to express different analyses of all-things-considered obligation, such as that for all reasons for $\neg\phi$, there is a (potentially complex) better reason, such and such, for ϕ, and thus that ϕ is all-things-considered obligatory.

Even without these extensions, in the present paper we developed a basic approach in the framework of justification logic to capture practical reasoning about all-things-considered obligations and *pro tanto* reasons.

References

[1] Alchourrón, C. E. and D. Makinson, *Hierarchies of regulations and their logic*, in: R. Hilpinen, editor, *New Studies in Deontic Logic*, Springer, 1981 pp. 125–148.

[2] Artemov, S., *Explicit Provability and Constructive Semantics*, Bulletin of Symbolic Logic **7** (2001), pp. 1–36.

[3] Artemov, S., *The Logic of Justification*, Review of Symbolic Logic (2008).

[4] Artemov, S., *Epistemic Modeling with Justifications* (2017).
URL `http://arxiv.org/abs/1703.07028v2`

[5] Artemov, S., *Justification Awareness*, Journal of Logic and Computation **30** (2020), pp. 1431–1446.

[6] Artemov, S. and M. Fitting, *Justification Logic*, in: E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, 2012, fall 2012 edition .

[7] Artemov, S. and M. Fitting, "Justification Logic: Reasoning with Reasons," Number 216 in Cambridge Tracts in Mathematics, Cambridge University Press, 2019.
URL `https://doi.org/10.1017/9781108348034`

[8] Bader, R., *Conditions, modifiers, and holism*, in: Lord and Maguire [31] .

[9] Boella, G. and L. van der Torre, *Institutions with a hierarchy of authorities in distributed dynamic environments*, Artificial Intelligence and Law **16** (2008), pp. 53–71.

[10] Chisholm, R., *The ethics of requirement*, American Philsophical Quarterly **1** (1964), pp. 147–153.

[11] Chisholm, R., *Practical reason and the logic of requirement*, in: S. Körner, editor, *Practical Reason*, Yale University Press, New Haven, 1974 pp. 1–17.

[12] Dalia, D., *Reasons have no weight*, Philosophical Quarterly **68** (2018), pp. 60–76.

[13] Dancy, J., "Ethics Without Principles," Oxford University Press, Oxford, 2004.

[14] Dietrich, F. and C. List, *A reason-based theory of rational choice*, Noûs **47** (2013), pp. 104–134.

[15] Dietrich, F. and C. List, *Choice-theoretic deontic logic* (2017), ms.

[16] Dietrich, F. and C. List, *What matters and how it matters: A choice-theoretic representation of moral theories*, The Philosophical Review (2017).

[17] Faroldi, F. L. G., "Hyperintensionality and Normativity," Springer, 2019, ms.

[18] Faroldi, F. L. G. and T. Protopopescu, *A Hyperintensional Logical Framework for Deontic Reasons*, Logic Journal of the IGPL **27** (2019), pp. 411–433.

[19] Fitting, M., *The Logic of Proofs, Semantically*, Annals of Pure and Applied Logic **132** (2005), pp. 1–25.

[20] Fitting, M., *A Quantified Logic of Evidence*, Annals of Pure and Applied Logic **152** (2008), pp. 67–83.

[21] Fitting, M., *Modal Logics, Justification Logics and Realization*, Annals of Pure and Applied Logic **167** (2016), pp. 615–648.
URL `http://melvinfitting.org/bookspapers/pdf/papers/ModalJustificationRealization.pdf`

[22] Goble, L., *Prima facie norms, normative conflicts, and dilemmas*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, College Publications, 2013 pp. 241–351.

[23] Hansen, J., *Deontic logics for prioritized imperatives*, Artif. Intell. Law **14** (2006), pp. 1–34.
URL `http://dx.doi.org/10.1007/s10506-005-5081-x`

[24] Hansen, J., *Prioritized conditional imperatives: Problems and a new proposal*, Autonomous Agents and Multi-Agent Systems **17** (2008), pp. 11–35.
URL `http://dx.doi.org/10.1007/s10458-007-9016-7`

[25] Horty, J. F., "Reasons as Defaults," Oxford University Press, 2012.

[26] Horty, J. F. and S. Nair, "The Logic of Reasons," forthcoming.

[27] Kearns, S. and D. Star, *Reasons as evidence*, in: R. Shafer-Landau, editor, *Oxford Studies in Metaethics*, Oxford Studies in Metaethics **4**, Oxford University Press, 2009 pp. 215–242.

[28] Kearns, S. and D. Star, *Weighing reasons*, Journal of Moral Philosophy **10** (2013), pp. 70–86.

[29] Kearns, S. and D. Star, *Weighing explanations*, in: I. Hirose and A. Reisner, editors, *Weighing and Reasoning. A Festschrift for John Broome*, Oxford University Press, forthcoming .

[30] Kuznets, R. and T. Studer, *Justifications, ontology, and conservativity*, in: T. Bolander, T. Braner, S. Ghilardi and L. Moss, editors, *Advances in Modal Logic*, Advances in Modal Logic **9**, College Publications, 2012 .

[31] Lord, E. and B. Maguire, editors, "Weighing Reasons," Oxford University Press, Oxford, forthcoming.

[32] Osherson, D. and S. Weinstein, *Preference based on reasons*, Review of Symbolic Logic **5** (2012), pp. 122–147.

[33] Osherson, D. and S. Weinstein, *Deontic modality based on preference* (2014), http://arxiv.org/abs/1409.0824.

[34] Parfit, D., "On What Matters," Oxford University Press, Oxford, 2011.

[35] Raz, J., "Practical Reason and Norms," Hutchinson, London, 1975.

[36] Raz, J., *Value and the weight of practical reasons*, in: Lord and Maguire [31] .

[37] Rohani, A., T. Studer and F. Faroldi, *Conditional obligations in justification logic* (2023), ms.

[38] Ross, W. D., "The Right and the Good," Oxford University Press, Oxford, 1930.

[39] Scanlon, T. M., "Being Realistic about Reasons," Oxford University Press, Oxford, 2014.

[40] Schroeder, M., "Slaves of the Passions," Oxford University Press, New York, 2007.

[41] Van De Putte, F. and C. Straßer, *A logic for prioritized normative reasoning*, Journal of Logic and Computation **23** (2013), pp. 563–583.
URL `+http://dx.doi.org/10.1093/logcom/exs008`

Deontic Logic in a Hierarchical Framework

Rodrigo Mena González [1]

Munich Center for Mathematical Philosophy (Ludwig Maximilians Universität)
Geschwister-Scholl-Platz 1
Munchen

Abstract

Systems of Deontic Logic usually ignore the fact that deontic propositions can be analysed in organisational contexts. They also ignore specific solutions to various logical problems, that emerge in such circumstances. To fill this gap, this paper proposes a class of formal framework to represent such contexts based on a general notion of 'hierarchy' as an ordered set of agents. Each agent has ascribed a set of variables representing states of affairs they are authorized (or may commit others) to produce. Consistent subsets of such variables are possible worlds for him and give place to traditional deontic operators. Two additional operators representing directive speech acts, aiming to reduce or widen his authority, generate new models each time. Thus, commands and obligations are considered independent from each other: one agent may be obliged to act even though no command has been uttered to do so.
This framework is flexible enough to distinguish different types of obligations on a purely formal ground. Individual and general obligations, but also contractual, democratic and customary obligations are explained in a similar fashion. Finally, natural solutions for deontic conflicts arise relying on the order of agents in every hierarchy as well as on prioritising hierarchies.

Keywords: Hierarchies, Deontic Logic, Deontic Conflicts.

1 Introduction

Most systems of Deontic Logic ignore the solutions arising from everyday law practice to conflicts among obligations. These solutions are well known for centuries of detailed study on real-world normative systems. Possible solutions to such conflicts in formal logic involve weakening the logic in order to allow up to some point inconsistencies among obligations without trivial results (as Da Costa and Carnielli suggest in [8]), the rejection of the deontic version of the Necessitation Rule (see [9] or [7]), or the development of non-monotonic systems: i) to acknowledge the existence of contradictory obligations and ii) to provide an escape route to such antinomies at the same time [12]. Paying

[1] rod.mena@campus.lmu.de; r.mena.g@gmail.com

tribute to both ideas seems to be one distinctive feature of almost all formal approaches to deontic conflicts.

As is usual in Logic, adding more structure to deontic models should provide one opportunity to deal with both requirements, but logicians have used this approach to study the connections between the notion of agency and that of obligation instead of normative antinomies. This is understandable considering that agency problems may be more pressing than mere contradictions from a philosophical point of view.

Hierarchies are maybe the most suitable type of structures to represent contradicting obligations and ways to avoid them, although it seems that these structures have not been studied in full detail, at least, in order to fulfil these goals. They have been studied concerning concepts such as power or authority [5], although as far as the author of the present paper knows, they do not coincide with the approach adopted here.

In the next sections, two different classes of models are described. The idea is to consider hierarchies as partially strict orders of agents (or sets of agents), each of which has attached a set of variables representing states of affairs open to be produced by them according to the hierarchy configuration (called its 'authority'). Two deontic operators recover Deontic Logic considering sets of such states. Later, two additional operators representing speech acts aimed to enlarge or reduce the authority of agents with some resemblance to those studied by Public Announcement Logic are analysed too. Different types of obligations are distinguished afterwards, and finally, well-known solutions coming from Law theory and practice are adapted to this general framework.

2 Hierarchical Models: The Basic Picture

The main goal of this paper is to extend to Deontic Logic some solutions that everyday legal practice use to avoid conflicts among obligations, adapting the formal language of Deontic Logic with only minor changes. So, let L be a language construed over the rule:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid \langle H \rangle_a \varphi \mid [H]_a \varphi \tag{1}$$

Every propositional variable such as $p, q...$ belongs to a set AT of atoms, and other connectives are defined as usual. Read the symbols $\langle H \rangle_a \varphi$ as "agent a is authorized to produce a state where φ is true in hierarchy H" and $[H]_a \varphi$ as "agent a ought to produce a state where φ is true in hierarchy H".

Let the frame $H = \langle A, <, AU \rangle$ be a hierarchy, with A as a set of agents, $<$ as a binary order relation between members of A, and AU as an authorization function.

A hierarchy H is a strict partial order. Call 'chain of command' to every path from a superordinate to subordinate agents that ends in a subordinate with no other subordinates. Call the node occupied by an agent as her 'position in the hierarchy', and let a function $f : A \mapsto \mathbb{N}$ determine the 'rank of a position', assigning a natural number starting with 1 to nodes at the top of the hierarchy and adding 1 to every immediate subordinate. According to common usage, consider that the biggest the rank of an agent, the closer its number approximates 1, so take $a < b$ if $f(a) < f(b)$. If an agent belongs to two or more different chains of commands in a hierarchy, f assigns to every position the biggest natural number deriving from all chains. Additionally, the *depth* of the hierarchy is the same as the rank of the last subordinate in its longest chain of command.

Take a hierarchy $H = \langle A, <, AU\rangle$ as example number 1, with $A = \{a, b, c, d\}$, $<= \{\langle a, c\rangle, \langle b, d\rangle, \langle c, d\rangle$. Ignore AU at the moment. It is possible to represent H as follows:

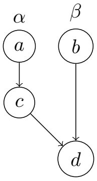


Fig. 1

There are two chains of command in this hierarchy: α and β. There is a label identifying both chains above their initial node. According to both chains, agents a and b have positions of rank 1. Agent c has a position of rank 2 and agent d has a position of rank 3 because 3 is the biggest rank that can be assigned to his position according to a chain to whom he belongs.

Let AU be a possibly partial *authorization function* that assigns to each variable in AT or its negation a set of members $a \in A$. Propositional variables or their negations in AU describe states that *may be* produced as a result of the actions of an agent. Let AU_a designate the inverse image of a under AU, and call it 'the authority of a'. When the context is clear enough, the subscript is omitted.

Take hierarchy H from example 1, with $AU(p) = \{a, b, c, d\}$, $AU(\neg p) = \{a, c\}$, $AU(q) = \{a, d\}$, $AU(\neg q) = \{d\}$, and $AU(r) = \{a, b, c, d\}$. It should look like this:

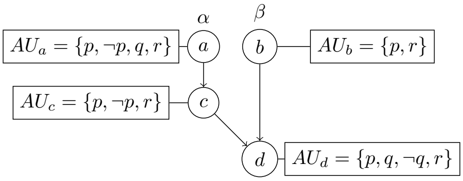

α
β
$AU_a = \{p, \neg p, q, r\}$
a
b
$AU_b = \{p, r\}$
$AU_c = \{p, \neg p, r\}$
c
d
$AU_d = \{p, q, \neg q, r\}$

Fig. 2

In modern bureaucracies (the epitome of a hierarchy), agents may be authorized to produce certain states of affairs but also to abstain from producing them. Thus, their authority may comprehend possible incompatible (inconsistent) results, being guaranteed the freedom to do or undo specific outcomes. In these cases, one agent could produce a state where p is true but also another where $\neg\, p$ is true. So, AU must be considered as a possibly partial function because we cannot assume that any agent not authorized to produce a state where p is true, is automatically authorized to perform actions resulting in states where $\neg p$ is true. Although the reason for this will become clear later, it is maybe important to say here that if an agent is only authorized to obtain a certain result but not its negation, he may be considered obliged to obtain it, so it seems unnatural to assume that every agent has always negative obligations if a propositional variable does not belong to her authority. Here, it is better to think that the agent cannot *intervene* in the final outcome neither to produce it nor to avoid it, being indifferent to him. Otherwise, it could be difficult to imagine models where agents lack any obligation at all, even though such models are perfectly conceivable. In any case, it is pretty normal to authorize the same agents to undo whatever they may be ordered to do. In fact, avoiding this kind of inconsistency in the distribution of authorizations is one way to solve some problems in deontic logic. However, imposing such a restriction on all models is not only inconvenient but unrealistic. It is worth noting that no hereditary condition is assumed among the AU sets of agents in the same chain of command.

Call MAX^a to every subset of AU_a comprising every atom in AU_a or its negation (if it also exists in AU_a), but not both. In a sense, all MAX^a are maximally consistent sets with respect to AU_a given that no two atoms p and $\neg p$ in AU_a belong to the same MAX^a, and that for every superset $\Gamma \subseteq AU_a$, Γ comprises at least one pair of atoms p and $\neg p$. The set $AU_a = \{p, \neg p, q, r\}$, in the example above, gives rise to two different MAX^a sets ($MAX_1^a = \{p, q, r\}$ and $MAX_2^a = \{\neg p, q, r\}$); the set $AU_b = \{p, r\}$, only one MAX set that is exactly the same AU_b, etc...

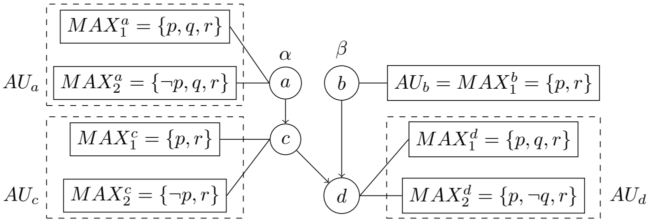


Fig. 3

Let $M = \langle H, V\rangle$ be a model construed over a hierarchy H, with V as a valuation function that assigns to A a set of propositional variables in AT, but not necessarily in AU_a for any a. Unlike AU, consider V as a total function.

While atoms in AU represent the authority of an agent (the range of actions an agent is authorized to do), atoms in the inverse image of A under V describe the actual world (called 'AW'), common for every agent in H.[2] To represent in the actual world the *fact* that agents have such and such power or obligation in the context of a hierarchy, it is necessary to attach special operators to the corresponding variable (or its negation) in AU_a. As has been said before, let $\langle H\rangle_a$ and $[H]_a$ be those operators.

So, define truth according to the following clauses:

- $M \vDash p$ if and only if $p \in AW$,
- $M \vDash \neg\varphi$ if and only if $M \nvDash \varphi$,
- $M \vDash \varphi \wedge \psi$ if and only if $M \vDash \varphi$ and $M \vDash \psi$,
- $M \vDash \langle H\rangle_a\varphi$ if and only if there is at least one $n \in \mathbb{N}$ such that $\varphi \in MAX^a_n$,
- $M \vDash [H]_a\varphi$ if and only if there is at least one $n \in \mathbb{N}$ such that $\varphi \in MAX^a_n$ and for all $n \in \mathbb{N}$ it is the case that $\varphi \in MAX^a_n$.

Definitions of validity and semantic consequence are as usual (a formula φ is valid here if and only if it is true in all hierarchical models, and it is a consequence of a set of formulas Γ if and only if is true whenever all formulas in Γ are also true).

The fourth clause says that an agent is able to produce a state where φ is true if there exists at least one consistent set of state descriptions she is authorized to obtain, containing φ. AU_a contains all the possible states authorized to a to produce, and every MAX^a is a maximally consistent subset of AU_a. It is convenient to think about AU_a as the set of all possible worlds immediately accessible to a, and every set MAX^a_n as one particular world in AU_a. Consequently, according to the fifth clause, whenever $\varphi \in MAX^a_n$ for all n, it is not only possible for a to produce φ, but also an obligation to do it. In other words, an agent is obliged to produce a given outcome just when the variable representing it *is* contained in AU_a, and it is not able to obtain a different incompatible state.[3]

Recall hierarchy H from the example, and let $V(p) = A$. The following figure represent the model $M = \langle H, V\rangle$

[2] As one of the referees correctly suggested, given that $V(A)$ assigns the same set of variables to all agents, an alternative way to define a model should be as a pair $\langle H, AW\rangle$, with $AU \subseteq AT$. The only reason to stick to our original definition is that V allows us to distinguish how different a description of the actual world and a normative one are.

[3] In order to validate D axiom ($[H]_a\varphi \rightarrow \langle H\rangle_a\varphi$), the existential requirement in the interpretation of the $[H]_a$-operator is unavoidable. According to this, another way to define the set of obligations of an agent is as the intersection of all possible worlds open for him in AU_a *whenever this intersection is not empty.*

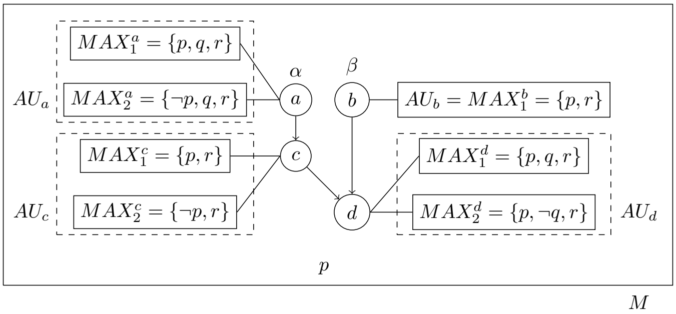


Fig. 4

Note that p, $\neg q$ and $\neg r$ are true in M. $\langle H\rangle_a q$ is also true because there is at least one MAX world accessible to a. $[H]_a q$ is true in M too, given that q is in all MAX worlds accessible to a. For the same reason, $[H]_n r$ is true for all agents $n \in A$. Finally, both $\langle H\rangle_c q$ and $\langle H\rangle_c \neg q$ are false given that neither q nor $\neg q$ belong to AU_c.

This way, all basic deontic operators are recovered here in a very weak logic, where both operators are not inter-definable with each other (at least in the usual way, '$[H]_a\varphi := \neg\langle H\rangle_a\neg\varphi$'), as a consequence of AU being a partial function.[4]

So, in the basic picture, it is possible to distinguish different types of obligations. Those binding only one agent are called 'individual' obligations ($[H]_a q$, in the example above); those binding sets of agents are 'collective' obligations ($[H]_n p$ for all agents $n \in B$ with B as the set of agents $B = \{b, d\}$). 'General' obligations are collective obligations forcing all members of the set A, to perform or to abstain to perform an action ($[H]_n$ for all $n \in A$).

3 Changes in the Simple Picture

3.1 Reducing AU_a

In the basic framework, agents cannot modify the set of powers and obligations they have. One way to represent these changes in a hierarchy is by adding special operators representing *commands* and *authorizations.*

The simplest way to interpret a command is as an act of forcing another agent (considered as an immediate successor in the hierarchy), to obtain a certain outcome. So, define an (immediate) successor function in the structure as a mapping $s : \wp A \mapsto \wp A$ assigning to every agent $a \in A$, one set of agents

[4] One of the referees of this paper correctly suggests that $[H]_a\varphi \rightarrow \langle H\rangle_a\varphi$ and $[H]_a\varphi \rightarrow \neg\langle H\rangle_a\neg\varphi$ hold in this system, but $\neg\langle H_a\rangle\varphi \rightarrow [H]_a\neg\varphi$ does not hold if AU_a is empty. In that case $\neg\langle H\rangle_a\varphi$ is true but $[H]_a\varphi$, false. In fact, as anyone could expect, the equivalence between $\langle H\rangle_a$ and $[H]_a$ only holds when AU_a is non-empty, but this is not a necessary feature of AU sets.

$B \subset A - \{a\}$ such that there is no agent $c \in A - \{a\}$ such that $a < c < b$ for every member $b \in B$.

No reasonable command may force an agent to produce states where p and $\neg p$ are true at the same time, but only when is forced to produce one of both options, the other turns out to be prohibited (that is, $\langle H\rangle_a\varphi \wedge \langle H\rangle_a\neg\varphi$ may hold for an agent a but if a is then forced to produce φ, $\neg\langle H\rangle_a\neg\varphi$ will be the case (or in other words, $[H]_a\varphi$). According to this, an agent is obliged to do something when being authorized to do so, is not authorized to abstain from doing it.[5] So, whenever an agent is forced to produce states where φ is true, his authority is temporarily cut preserving only those MAX subsets of AU_a containing φ.

So, define the truth of a new command !-operator according to clause (!1) as follows:

(!1): *$M \vDash !_{ab}\varphi$ if and only if $M \vDash \langle H\rangle_a\varphi$ and b is such that $b \neq a$, and it is the case that $s(a) = b$, that $M \vDash \langle H\rangle_b\varphi$, and that $M' \vDash [H']_b\varphi$.*

Subscripts are added to the operator to make clear the issuer and the receiver of the command, otherwise, it would be not clear who of two different superordinate agents oblige a common subordinate to do something. Without subscripts, the command could be true for all of them. Additionally, the apostrophe in M and H indicates the new structure arising after the command issued by a superordinate agent. H' is the new hierarchy arising from a true command. As such, it is a modification of the original hierarchy H. In H', the new function AU' remains exactly the same with respect to the original function AU except that according to AU', AU_b lacks $\neg\varphi$ if this atom already belonged to it. So, $AU'(\neg\varphi) = \{a \in A : \neg\varphi \in AU_a\} - \{b\}$. Consequently, $H' = \langle A, <, AU'\rangle$, and M' is the new model that comprises $\langle H', V\rangle$.

In any case, according to this clause, one can consider that an agent a has committed another agent b to perform φ only when:

(i) a has the authority to perform an action φ,

(ii) there is a subordinate b that is a successor of a (call this as the 'order condition' or OC),

(iii) b is also authorized to perform φ (that is, $\varphi \in AU_b$ holds for b),

(iv) the authority of b is reduced to φ, verifying $[H']_b\varphi$, after the command.

Not every agent in a hierarchy is entitled to command others to produce a specific outcome, nor is entitled to command others to produce any outcome the agent wants. It is natural to think that an agent can only issue orders within their own scope of authority (requirement 1).[6] No one expects the Secretary

[5] Coincidently, a very similar definition of an obligation, based on the idea of normative ability can be found in [15, p. 397 and 398].

[6] This does not necessarily mean that for every agent a and a subordinate b, the formula $\langle H\rangle_a\varphi \rightarrow \langle H\rangle_a!_{ab}\varphi$ always hold. It will only hold for an agent a if !-formulas are previously allowed to belong to AU_a sets, and after that, when $!_{ab}\varphi \in_{AU_a}$ also holds for a.

of Defense to issue rules about Public Health. It is pretty clear also, that the only states of affairs an issuer of a command can force others to produce are those within the authority of the receiver (requirement 3). A captain cannot order his lieutenants to act beyond the call of duty. And finally, it is difficult to consider that an agent has been commanded to do something if the agent is already obliged to do so. In these cases fail a preparatory condition of the speech act.

It is interesting to point out here that the last clause defines the idea of hierarchical obligation in a way independent of the act of commanding something. In some normative systems, it is commonly accepted that the obligation of an agent emerges at the same time as the correlative command of another agent, who forces the first to give or do whatever the second agent may require from him. In this setting, one must accept that in a hierarchy commands and obligations are not necessarily symmetrical, and that forcing anyone to do something that the agent already has to do changes nothing.

In this setting, one superordinate can only bind immediate subordinates to produce some state, because the idea here is to simplify the description of hierarchies. It is worth noting that, according to the last clause agents of rank 1 cannot be forced by anyone in the hierarchy, and agents of the last rank cannot force anyone else to produce a result true in some states of affairs. So, they don't have any other choice but to do the required action by themselves.

So, let $!_{ac}p$ hold in M from example 1. This gives birth to the following model M':

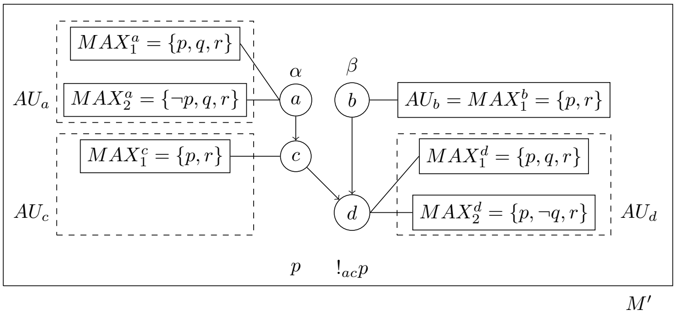


Fig. 5

Accordingly, MAX_2^c disappears, turning MAX_1^c be the only ideal world accessible to c. Note that a formula $!_{ac}q$ should not be true because $q \notin AU_c$. This is consistent with the intuitive idea that it is not reasonable to force an agent to obtain a state of affairs whose production is not in his authority, nor to disqualify the agent to obtain a state of affairs which he could not produce in the first place.

It is possible to generalize the interpretation of the !-operator if one drops the restriction posed on commands, to be directed only to immediate subordinates. Adding subscripts to commands indicating the recipient of them in each case is enough to open a door to a more fine-grained analysis of norms. Let $!_{ab}\varphi$ be a command directed from agent a to b. The clause for such a formula would be, then:

(!2): *$M \vDash !_{ab}\varphi$ if and only if $M \vDash \langle H \rangle_a \varphi$ and there is an agent $b \neq a$, such that $a < b$, it is the case that $M \vDash \langle H \rangle_b \varphi$, and it is the case that $M' \vDash [H']_b \varphi$.*

If the last clause is admitted, letting the formula $!_{ad}\neg q$ hold in M' give birth to the following model M'':

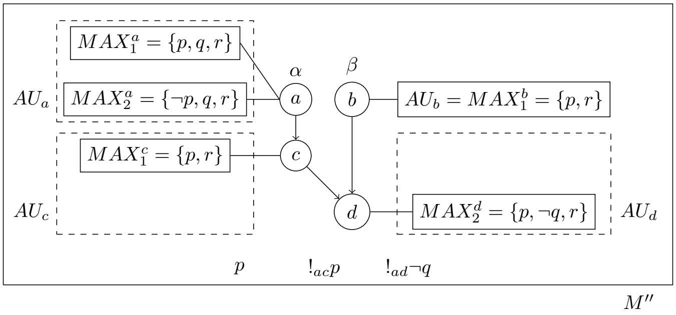


Fig. 6

Up to now, a superordinate agent cannot eliminate an obligation forcing a subordinate. One way to do it should be allowing superordinate agents to make authorizations, turning obligations into permissions (this alternative is to be explored in the next section.) A second way to do it should be modifying the !-clause in a specific way in order to allow agents in a higher position to eliminate variables from AU_a sets of agents in a lower position, whose negation is already not in AU_a:

(!3): *$M \vDash !_{ab}\varphi$ if and only if $M \vDash \langle H \rangle_a \varphi$ and there is an agent $b \neq a$, such that $a < b$, and it is the case that $M' \nvDash \langle H' \rangle_b \psi$, with ψ as the atom $\neg p$ if $\varphi = p$, or the atom p if $\varphi = \neg p$.*

Now, it is not even necessary that the receiver of the command be previously authorized to obtain the state of affairs commanded by his superior. It is enough that the agent in a superordinate position commands something that eliminates an incompatible state of affairs from the AU_a set of his subordinates. Thanks to this clause, the formula $!_{ac}\neg p$ added to model M'', for example, gives place to the model M''':

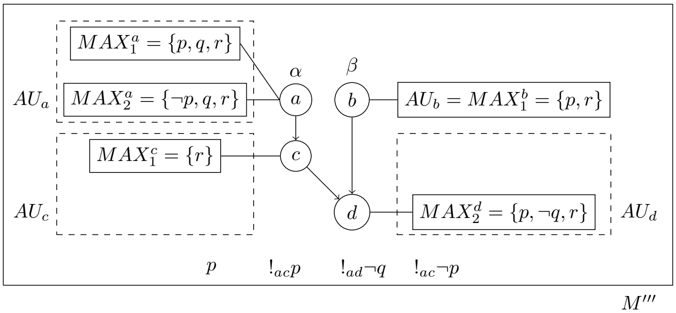


Fig. 7

3.2 Increasing AU_a

Up to now, the only normative action one agent could perform was to commit someone to produce some state of affairs. This way, one agent reduces the authority of another. But there are at least two other different possible available normative actions: to increase the authority of an agent (sets of agents), and to change the position of an agent (sets of agents) in the hierarchy. Only the first alternative will be analysed here. So, this requires the addition of a new authorization +-operator in our language, which should be interpreted as follows:

(+1): *$M \vDash +_{ab}\varphi$ if and only if $M \vDash \langle H \rangle_a \varphi$, there is an agent $b \neq a$, such that $s(a) = b$, $M \nvDash \langle H \rangle_b \varphi$, and it is the case that $M' \vDash \langle H \rangle_b \varphi$.*

Again, the apostrophe indicates a new hierarchy, where AU_b has changed: the utterance of a formula $+_b\varphi$ by an agent a gives place to a new model $M' = \langle H', V \rangle$ with V as always, but with $H' = \langle A, <, AU' \rangle$. A and $<$ remain the same, but $AU'(\varphi) = \{a \in A : \varphi \in AU_a\} \cup \{b\}$.[7] Also, as happened with the !-operator, it seems that four conditions must be met to consider an authorization true (or at least, effective):

(i) the agent a that issued an authorization has the authority to produce states of affairs where φ is true,

(ii) there is a subordinate b that is a successor of a,

(iii) b is not authorized to perform φ (that is, $\varphi \notin AU_b$),

(iv) the authority of b includes φ after the command, and because of that the new set $AU'_b = AU_b \cup \varphi$ is in H'.

Here it is also natural to think that agents can only be authorized to obtain certain states of affairs that lie within the scope of authority of the authorizing

[7] The referee correctly pointed out that here it is implicit that AU remains the same for all other formulas.

agent (requirement 1). It would be surprising that the Secretary of Defense authorize hospitals to apply a special vaccine or a medical procedure to everyone. One should expect the Secretary of Defense to authorize others to produce states he could produce by himself. On the other hand, it is not reasonable to consider that one agent has been authorized to obtain states he was already able to produce before the authorization (requirement 3). The preparatory conditions of such authorization would fail as happened to the !-operator, turning this speech act useless.

It is worth noting that increasing the authority of an agent a doesn't mean that a necessarily acquires the right to perform or to omit an action. If an agent is only granted to produce a state where φ is true not being also authorized to produce $\neg\varphi$, then the authorization looks very similar to a command, in fact, it is equivalent to a command to produce a state the agent cannot produce before. Here 'authorization' has been used for lack of a better word. So, one authorization grants the right to produce or omit the production of a state only when its negation has previously been granted to the agent.

Take model M''' and let $+_{ac}p$ be true in it. The resulting model M'''' should be identical to M''. According to M'''', agent c is now obliged to obtain a state of affairs where p and r hold. This is an example of an authorization that is equivalent to a command from a to c as 'you must produce a state of affairs such that $p \wedge r$'.

The interpretation of the +-operator may be generalized in the same way the !-operator was in the previous section:

(+2): *$M \vDash +_{ab}\varphi$ if and only if $M \vDash \langle H \rangle_a\varphi$, and there is an agent $b \neq a$, such that $a < b$, $M \nvDash \langle H \rangle_b\varphi$, and it is the case that $M' \vDash \langle H' \rangle_b\varphi$.*

Thanks to this clause, if $+_{ad}q$ is added to model M'''', the resulting model M''''' will be exactly as M', being $\langle H \rangle_d q$ and $\langle H \rangle_d \neg q$ true again.

One last remark must be added now. If it is allowed to iterate !- and +-operators and to include ! or +-formulas in AU-sets, it is possible to represent more changes in the authority of agents. It could be possible for a superordinate agent to grant the subordinate agents the authority to perform normative actions. Thanks to this modification, the first agent could authorize others to address or being able to receive new commands. A formula like $!_{ab}\,!_{bc}\varphi$ is true for agents $a < b < c$, for example, if only $!_{bc}\varphi \in AU'_b$ holds for b, and consequently, $[A]_b!_{bc}\varphi$ also holds. This means that b ought to command c to obtain φ. This happens when a General says to a Colonel: 'order your men to attack!', in the face of the enemy. It is easy to think of the truth conditions and possible examples of formulas built by all other possible combinations of ! or +-operators. [8]

[8] There is an obvious similarity among formulas $[H]_a!ab\varphi$ and $[H]_a + ab\varphi$, and those formalizing the theory of normative position in the style of Kanger-Lindahl (KL) works, although this framework was not intended to fulfil such goal. The 'position of an agent in a hierarchy' has a different meaning here but, due to its generality, it should be also possible to distinguish here relevant deontic positions in the sense of KL theory.

4 Relations with the Standard System of Deontic Logic (SDL).

The hierarchical framework described here (HF) provides a more general approach to Deontic Logic than the Standard System (SDL). SDL rests on at least three assumptions that HF lacks: SDL presupposes the normative validity of deontic formulas (those with deontic operators), their completeness and consistency. The first is very subtle, the other two are more easy to grasp.

According to a descriptive interpretation of Deontic Logic, the validity of a formula with deontic operators depends on the existence of a normative system, the utterance of a corresponding norm, and the fact that the utterer is competent to issue such a norm. As Kelsen elegantly puts it, the validity of a norm is the particular mode of its existence [10, p. 10]. And this coincides, partly, with the concrete act of will that gives birth to it. Under this kind of interpretation, the validity conditions of a norm turn out to be the same as the truth conditions of the proposition describing it. But the resulting logic cannot coincide with SDL. Many of its theorems become automatically invalid. On the other hand, if one sticks to a normative interpretation to preserve SDL, deontic formulas should be considered without regard to any normative system, nor to the competence of any issuer or receiver. But now the validity of deontic formulas is up to some point 'presupposed' [9, p. 30 and 31]. In fact, according to the usual semantics of SDL, a formula Op is true or false only depending on whether p is true in every ideal world accessible from the actual world, in a way completely independent of the grounds that make it obligatory. In other words, the normative status of ideal worlds (their *ideality*) is already given, and there is no account of how this happens.

HF does not assume the validity of deontic formulas. As on the descriptive interpretation above, the validity of normative formulas depends explicitly on the membership of an agent to a normative system (a hierarchy), and his competence according to it (the AU set attached to him). In a more general

The positions agents occupy in a hierarchy are defined in a relational way based on the existence of an order among them. This order justifies the power of one agent over his subordinates. All these elements are absent from the usual theory of normative positions. According to Sergot [14, p. 357], KL theory only has a deontic logic component, an action logic component, and a method for generating all logically possible normative relations between two agents. So, even though KL also defines positions on relational grounds it lacks all the structure that hierarchical frameworks rely on and, especially, the power dimension that also justifies the normative force of deontic operators here.
The deontic logic component of the KL theory is the Standard System of Deontic Logic, so any theory of normative positions built over hierarchical frameworks will differ from KL in all points described in the next section. The action logic component in KL theory is different from those in the class of hierarchical frameworks too. KL theory uses the STIT theory approach, focused on the end result of actions than on state changes [14, p. 359]. Hierarchical Frameworks use the two new ! and +-operators that increase or decrease the authority of an agent instead. They are not strictly related to actions but to changes in the legal capacity of agents. This fact poses an important restriction on a theory of normative positions built on a hierarchical framework because this could only refer to what has been called as *normative actions*, like commands and authorizations.

sense than Kelsen's, the membership of an agent to a hierarchy determines its 'imputability'. Obligations emerging from a hierarchy are not imputable to agents that do not belong to it. Conversely, the production of a state of affairs that is not mandatory nor permitted in the hierarchy is not imputable to its members, being indifferent to them. The ideality of MAX sets follows from the specific role they play in these ordered structures of agents. Their content is determined by the agents in a higher position on the hierarchy, and its normative force derives from this order. From a strictly formal point of view, it is not necessary to commit oneself to the social nature of such an order but to acknowledge its existence and provide a way to model it. Consequently, HF gives place to an openly different weaker logic, closer to Alchourrón's logic of normative propositions than to a proper Deontic Logic [1].

Alchourrón [1, p. 264] shows that a logic based on a descriptive interpretation of norms (a logic of normative propositions) is isomorphic to SDL when the first admits two more assumptions the second tacitly accepts: normative completeness and consistency. A concrete normative system is complete if it lacks gaps, meaning that every possible state of affairs has a normative status: it is permitted or prohibited, and therefore, the formula $Pp \vee \neg Pp$ is true for every variable p in the language ([1, p. 259]; [6, p. 137]). That formula is, as expected, a theorem of SDL and a modal instance of the law of excluded middle. In the ideal-worlds semantics of SDL, this is a consequence of a total valuation function assigning variables to all possible worlds. So, if a variable is not assigned to a world, its negation is. On the contrary, HF rejects normative completeness. The formula $\langle H\rangle_a\varphi \vee \neg\langle H\rangle_a\varphi$ is false whenever neither φ nor $\neg\varphi$ belongs to AU_a and, therefore, to any MAX^a set for any agent a. This is a consequence of being AU a partial valuation function, so it is not admissible to assume that $\neg p \in AU_a$ for any a whenever $p \notin AU_a$. For the same reason, it is not admissible to consider the falsity of a formula $[H]\neg p$ as a weak permission of p. If $p/inAU_a$ for some a, the falsity of $[H]_a\neg p$ means that $\langle H\rangle_a p$ is true to a, but if $p/notinAU_a$, then p is normative indifferent for a and, therefore, not imputable to him.

A normative system is consistent when forces its agents to obtain only mutually compatible states of affairs, so $\neg(Pp \wedge \neg Pp)$ is true for all p in its formal language. That formula is also a theorem of SDL, but only some of its instances hold in HF. The restriction posed over MAX sets to be maximally consistent is responsible in HF for every hierarchy to be consistent but this is not granted for agents belonging to two different hierarchies. Take for example two hierarchies $H = \langle A^H, <^H, AU^H\rangle$ and $I = \langle A^I, <^I, AU^I\rangle$ and an agent a such that $a \in AU^H$ and $a \in AU^I$. Let $AU^H_a = p$ and $AU^I_a = \neg p$. So, $[H]_a p$ is true to H and $[I]_a\neg p$ true to I, given that all MAX^a in H contain p and likewise all MAX^a in I contain $\neg p$. Consequently $[H]_a p \wedge [I]_a\neg p$ is true according to HF, but neither $[H]_a p \wedge [H]_a\neg p$ nor $[I]_a p \wedge [I]_a\neg p$ can be true because it is not possible that p and $\neg p$ hold at the same time for every MAX sets of a in H or I.

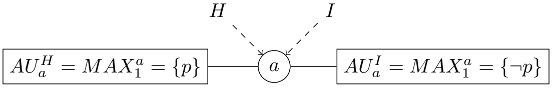


Fig. 8

5 Hierarchies based on the power set of agents.

5.1 A new picture

Slightly different modifications of the basic framework could provide a deeper understanding of normative systems, beyond the scope of usual Deontic Logics, if hierarchies are also defined as strict orders of subsets from the power set of A instead of just the set A itself.

Take a hierarchy as a structure $H = \langle \mathbb{A}, <, AU \rangle$, with $\mathbb{A} \subseteq \wp A$. This way, collective obligations may receive a very natural representation. In the basic picture, the position of every agent must be specified individually. There is no way to represent in a single position a group of agents even if their AU_a sets contain the exact same set of atoms. This is not exactly a problem in the basic picture but poses many limitations to representing commands that can emerge in some concrete hierarchies.

Take the following hierarchy $J = \langle A, <, AU \rangle$, with $A = \{a, b, c, d, e\}$, and $<= \{\langle a, c\rangle, \langle b, c\rangle, \langle c, d\rangle, \langle c, e\rangle\}$. If one wants to make both agents in the highest positions and both agents in the lowest position to work as a collective, it may be necessary to let $AU_a = AU_b$ and $AU_d = AU_e$. In any case, the only way to represent the resulting hierarchy according to the basic picture is like in the picture at the left. Hierarchy J should be depicted as in the right, nonetheless.

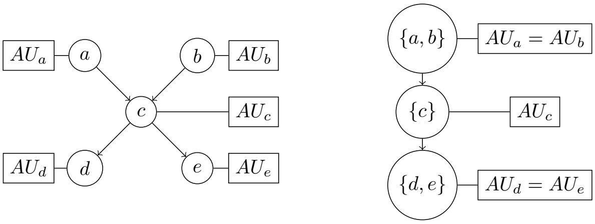


Fig. 9

In order to represent J in HF, it is inescapable to admit nodes to be occupied by sets of agents instead of only single agents, like in the picture on the right. If this is accepted, it could be possible to find agents that play two roles in a single hierarchy, like in the next hierarchy K:

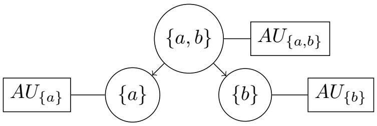


Fig. 10

This new feature may be sometimes inappropriate to reflect concrete hierarchies, so it could be possible to impose one additional 'disjoint condition' to prevent the possibility that one agent is in two different positions at the same time:

(DC) *Every position in a hierarchy must be occupied by disjoint subsets from the power set of agents.*

If the 'disjoint condition' is dropped, HF can now represent very complex types of organizations and equally complex types of obligations. Now it is perfectly possible that a subordinate $b > a$ be a member of a set of agents that commits a to produce a desired outcome, as happens in the following picture. So, consider every set of agents as a special entity different from its members. So, by dropping DC it is possible to represent commands that invert the order relation of the hierarchy. Take as an example hierarchy L:

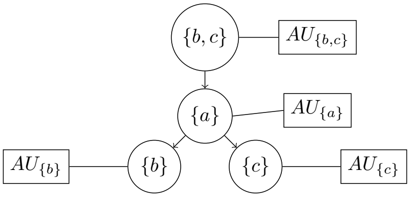


Fig. 11

By dropping DC, many types of hierarchies emerge. Hierarchies of singletons are still possible, as well as disjoint hierarchies, but now it would be possible to identify *contractual hierarchies*, as those where members of sets in a subordinate position also belong to sets on a superordinate position, like in the last two figures (hierarchies K and L). A coherent picture of the binding force of a contract in HF requires the set of its parties to be considered as superordinate to each of them in a new hierarchy born out of their common will. In the hierarchy K of Figure 10, agents a and b may represent the parties. Their obligations arise from a special entity formed by their agreement, portrayed by $\{a, b\}$ (called 'the contract').

The set of all parties must possess full authority to produce all the results that every party is obliged to obtain according to a contract, even though each party may have only a part of such authority. In other words, the authority of the set of parties is the union of every portion of their individual authority they choose to contribute to the contract. In Figure 10, this means that $AU_{\{a,b\}} \subseteq$

$AU_a \cup AU_b$ must hold in K. The normative ability that arises to each of them because of the contract is also limited to such a union set, obviously.

Finally, it should also be noted that it is possible to construct new hierarchies as the union of two or more other hierarchies. Here, for example, two hierarchies $H = \{A^H, <^H, AU^H\}$ and $I = \{A^I, <^I, AU^I\}$ can give place to a new hierarchy $J = \{A^{H\cup I}, <^{H\cup I}, AU^{H\cup I}\}$. It is possible to construct the product of hierarchies using the intersection operation on the sets of agents, order relations and authority sets of two or more prior hierarchies. This way, it would be possible to identify the sum and product of hierarchies as well.

5.2 Types of Obligations in this New Picture

As has been said before, contractual obligations should be defined here as deriving from the command of the very set of agents to whom the obliged agents belong. Many other obligations may be also analysed in this framework: individual and general obligations, and among these, also democratic and customary obligations. It is possible to explain the last two obligations with regard to purchase contracts.

According to an old theory from Civil Law, in these contracts, one party must give the purchased thing (the 'seller'), to another which must pay the price (the 'buyer'), and all other agents not in the contract must refrain to disturb the ownership rights the second agent acquired over the thing.

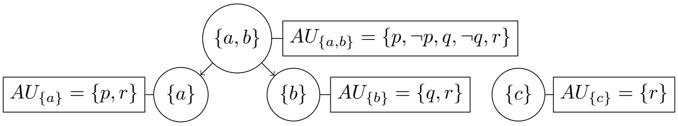


Fig. 12

Figure 12 shows a hierarchy $J = H \cup I$ with H as the purchase contract binding the buyer'a' and the seller 'b', and I as a hierarchy formed by one single agent c, who is a third person outside the contract. Let 'p' means 'agent a pays the price'; 'q', 'agent b gives the thing', and 'r', 'the agent refrains himself of disturbing ownership rights of others'. The binding force that obliges agents that are not a party in the contract (the truth of the formula '$[J]_c r$') doesn't emerge from the contract but should be considered in it. It is not possible for the members of the set of parties (the set $\{a, b\}$) to command agents outside the contract to do anything because these are not subordinate agents of the former. This means that the obligation of the agents outside the contract is true because of another reason. There are two possible answers here: considering general obligations as *democratic* or as *customary* obligations.

A 'democratic' obligation is just a contractual obligation issued by the set of all singletons in a hierarchy (maybe according to an aggregation rule such as the majority rule) that forces one or more agents in the hierarchy to do something. It must be singletons to exclude groups to be considered as agents in a democracy.[9] In other words, a democratic obligation arises from a democratic command. In these cases, the complete picture should require a set of all members of singletons ($\{a, b, c\}$ in the following picture) and a democratic command forcing everyone ($!_{\{a,b,c\}\{a\}}r$, $!_{\{a,b,c\}\{b\}}r$, and $!_{\{a,b,c\}\{c\}}r$), not to disturb the property rights of other agents, forcing the hierarchy created by the purchase contract to be part of a bigger hierarchy $L = H \cup K$ with J as the purchase contract, and $K = \langle A, <, AU\rangle$ as a democratic hierarchy created by all singletons $A = \{\{a\}, \{b\}, \{c\}\}$, just like the one in figure 13:

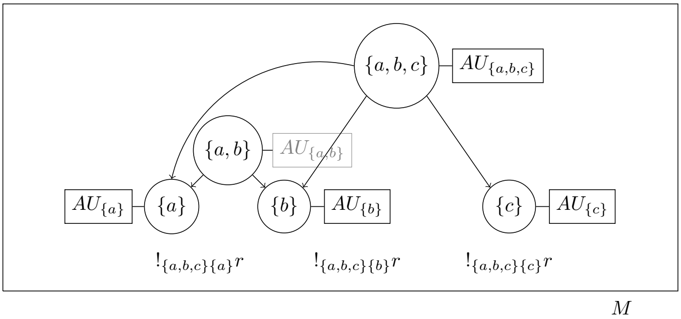


Fig. 13

A 'customary obligation' is one that forces the entire set of agents in the hierarchy but which was addressed by no agents.[10] One can think about these obligations as being set by default. The difference between what we call 'democratic' and 'customary' obligation lies here in which is the author of every kind: the author of the first one is the set of all agents A and the author of the last one, the empty set. Moral obligations sometimes may be considered customary obligations, sometimes not (i.e. when they are issued by a specific agent namely a prophet or a priest). So, if the general obligation r is interpreted as a customary obligation with respect to the purchase contract of the example above, the hierarchy of the model of the whole situation should coincide with Figure 12.

[9] The difference between individual and general obligations depends on the proportion of agents obliged in a given situation. That is why it is not necessary to use different operators for all these different types of obligations in this framework: a simple verification is enough to identify them depending on the number of agents obliged in each case.

[10] This concept tries to capture the legal notion of customary obligation, as a 'practice repeated for a long time and generally accepted as having acquired the force of law'[13, p. 56]. As a common practice, has no specific author, but it is believed it is obligatory for all agents. This way, this is a simpler concept of customary obligation than that of Bicchieri in [3].

6 Old Solutions with their own problems.

HF provides some solutions to everyday problems in normative systems. One obvious problem is that two norms may contradict each other. This could happen, at least, when one of the following pairs of formulas holds in a model [4, p. 187]:

- $[H]_a\varphi$ and $[I]_b\neg\varphi$,
- $[H]_a\varphi$ and $\langle I\rangle_b\neg\varphi$,
- $[H]_a\neg\varphi$ and $\langle I\rangle_b\varphi$,

for any hierarchies H and I, and any agents a and b. So, if one takes α and β to represent the formulas of these pairs, they may give place to the following types of conflicts in HF.

(i) α and β hold for a single agent a in a hierarchy H;

(ii) α and β hold for different agents a and b in a hierarchy H;

(iii) α and β hold for a single agent a in two hierarchies H and I;

(iv) α and β hold for different agents in different hierarchies H and I.

Type-i conflicts may be divided into three sub-types: the issuer of a command or authorization is a single agent in a higher position (type i-a), the issuers are two agents in a higher position but in the same chain of command (type i-b), the issuers are two agents in a higher position but in different chains of command (type i-c). Figure 8 depicts type iii conflicts.

Many different solutions to solve these problems may be found in the literature, ignoring that an obligation may be analysed in the context of a hierarchy. So, solutions based on paraconsistent and defeasible deontic logics have been presented elsewhere. Here, a set of traditional solutions from normative systems will be presented.

One obvious solution is to distribute among all agents only consistent sets of authorizations. But this is a very unrealistic solution, so it is better to accept that no matter what, conflicting commands and obligations can always appear. Standard solutions in everyday law practice admit the possibility of conflicts and depend on the order of issuers of a command or authorization (recall that an obligation may arise from one or the other), or on an order of hierarchies.

Type i-b conflicts are automatically solved in HF considering that a true command at a superior level could force all his subordinates to fulfil his orders if (!1) clause is accepted. Specific rules are generally added if clause (!2) is accepted instead, stating that commands from an agent in a higher position in the hierarchy supersede commands from a lesser agent in the hierarchy (as the classical aphorism *lex superiori derogat lex inferiori* suggests with respect to law conflicts):

(LSDLI) "*Whenever* $!_{ac}\varphi \to [H]_c\varphi$ *and* $!_{ac}\varphi$ *hold, but also hold* $!_{bc}\neg\varphi \to [H]_c\neg\varphi$ *and* $!_{bc}\neg\varphi$*, it must be case that* $[H]_c\varphi$ *if* $f(a) < f(b)$".

This solution could also be applicable to type i-c conflicts. Recall that function

f assigns a rank to the position of every agent in the hierarchy.

One general solution in HF that may solve all problems of type-i and ii is to consider that one first command crop the authority of a subordinate agent turning ineffective whatever a second incompatible command could intend from him, as in a first come first serve situation. A second command forcing him to produce an incompatible state of affairs will not automatically restore his previous authority nor change it to authorize him to do the opposite. Conflicting commands may eliminate all variables in the AU_a set of an agent a, if (!3) clause is admitted in a model, turning every state of affairs indifferent for a, as happens in the neutral view of normative conflicts studied in [11]. But this is opposed to the usual practice of law repealing, according to which a posterior command issued by an agent of the same rank or higher should derogate a prior one (summarized by old Baldo de Ubaldi's aphorism '*lex posterior derogat legi priori*'). The *lex posterior* solution is nevertheless possible in these frameworks provided the second order enlarges the previously reduced authority, and subsequently cut it again, to force the agent to obtain the opposite incompatible desired state. So, instead of simply issuing a $!_{bc}\neg\varphi$ order, the second superordinate agent $b < c$ should also utter an authorization $+_{bc}\neg\varphi$ as a previous step in the same normative action.

Type-i and ii conflicts may also arise in an indirect way. The command or authorization issued by one or two different agents in a higher position should derogate all other obligations that imply the first by contraposition. These cases are usually called 'implied derogations', and correspond up to some point to the *lex specialis derogat legi generali* solution.

Finally, to solve conflicts of types iii and iv it is still possible to add another condition such as

[PH] *"Whenever* $!_{ac}\varphi \rightarrow [H]_c\varphi$ *and* $!_{ac}\varphi$ *hold, but also* $!_{bc}\neg\varphi \rightarrow [B]_c\neg\varphi$ *and* $!_{bc}\neg\varphi$*, it must be case that* $[H]_c\varphi$ *if* $A \lhd B$*"*,

being '$\lhd$' a symbol stating a priority order among obligations depending on which hierarchy they belong (a priority of hierarchies, PH), such that $A \lhd B$ holds whenever a hierarchy A must be considered more important than another hierarchy B. This is a different, simpler approach to hierarchies of regulations than that Alchourrón and Makinson presented in [2], and should not be considered definitive. As Alchourrón and Makinson said, there is no obvious solution to the problem of prioritising obligations and the solution adopted here is in no way easy to determine. *Lex superiori* and *lex posteriori* solutions have been proposed here as naturally deriving from the hierarchy of agents that commit subordinates to do or omit something, by means of their commands and authorizations, and not as a relation that emerges directly among obligations as such. Furthermore, many different criteria may be invoked to define an order among hierarchies and it is not clear whether a formal approach to deontic logic favours one among all others or not.

Anyway, this is not over. All these solutions may oppose each other in some kind of second-level deontic conflict [4, p. 217]. For example, which solution is applicable if a posterior order is opposed to an order issued by an agent of a superior rank? The traditional way to solve this problem is to limit one of these rules (the *lex posterior* solution) to apply only in case of two opposed superordinate agents have the same rank. That is, to consider the *lex superiori* as a stronger criterion.

7 Conclusion

Hierarchical frameworks are very flexible and provide solutions to deontic conflicts in a natural way. These solutions are well-known in everyday law practice but are usually ignored from a formal point of view. They also provide the opportunity to model complex normative facts that the standard system can't, thanks to different types of obligations that may be distinguished here.

It is maybe true that HF oversimplify the way real-world normative systems work, but they have not been designed to provide a full picture of such systems, nor solutions to all open problems in Deontic Logic. It is pretty clear that many problems from Deontic Logic remain unsolvable under this framework. Although may not be of central interest, it seems that problems related to $O\top$ have a solution here, due to the weakening of the necessitation rule, derived from the fact that AU is only a partial function. Considering deontic modalities not to range over maximally consistent subsets of the authority of every agent but to the whole inconsistent set may provide another solution here. This is going to be studied in another paper.

Impossible norms and Kant's Law are problematic here also. Kant's Law states that 'anything morally obligatory for an agent must be *within the agent's ability*'[9, p. 67]. In the hierarchical approach, the factual ability is ignored and the notion of 'authorization' (as a deontic ability) is considered instead. The definition of ability in terms of the possibility modality remains available and nothing prevents us to consider the set of two formulas $[H]_a p$ and $\neg\Diamond_a p$ consistent (for any hierarchy H), and therefore, after the addition of a theorem such as $[H]_a p \rightarrow \Diamond_a p$ (a possible definition of factual ability on Deontic Logic), derive also $\Diamond_a p$ by *modus ponens.*

Although it seems adequate to provide a frame for the normative positions theory, this is also a pending task concerning HF.

In any case, there is no reason to deny hierarchical frameworks the possibility of being adapted to provide solutions to all these problems as much as other representations of normative systems are adaptable in such terms too. The solutions HF provide for normative conflicts at least, seem very simple and endorsed by practice.

References

[1] Alchourrón, C., *Logic of norms and logic of normative propositions*, Logique et Analyse **12** (1969), pp. 242–268.

[2] Alchourrón, C. and D. Makinson, *Hierarchies of regulation and their logic*, in: R. Hilpinen, editor, *New Studies in Deontic Logic: Norms, Actions, and the Foundations*

of Ethics (1981).

[3] Bicchieri, C., “Norms in the Wild,” Oxford University Press, New York, 2017, 1 edition.

[4] Bobbio, N., “Teoría General del Derecho,” Editorial Temis, Colombia, 2002, 2 edition.

[5] Bochenski, J., “¿Qué es Autoridad?” Herder, Barcelona, 1979, 1 edition.

[6] Bulygin, E., *Lógica deóntica*, in: M. J. Alchourrón, Carlos and R. Orayen, editors, *Lógica. Enciclopedia Iberoamericana de Filosofía* (2013).

[7] Chellas, B., “Modal Logic: An Introduction,” Cambridge University Press, Cambridge, 1980, 1 edition.

[8] Da Costa, N. and W. Carnielli, *On paraconsistent deontic logic*, Philosophia **16** (1986), pp. 293–305.

[9] Hilpinen, R. and P. McNamara, *Deontic logic: A historical survey and introduction*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems* (2013).

[10] Kelsen, H., “Pure Theory of Law,” The Lawbook Exchange Ltd., New Jersey, 2008, 1 edition.

[11] Kulicki, P. and R. Trypuz, *Multivalued logics for conflicting norms*, in: O. Roy, A. Tamminga and M. Willer, editors, *Deontic Logic and Normative Systems. 13th International Conference, DEON 2016, Bayreuth, Germany, July 18-21, 2016* (2013).

[12] Nute, D. and X. Yu, *Introduction*, in: D. Nute, editor, *Defeasible Deontic Logic* (1997).

[13] Saint Dahl, H., “McGraw-Hill’s Spanish and English Legal Dictionary,” McGraw-Hill, New York, 2004, 1 edition.

[14] Sergot, M., *Normative positions*, in: H. J. P. X. v. d. M. R. Gabbay, Dov and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems* (2013).

[15] Wooldridge, M. and W. van der Hoek, *On obligations and normative ability: Towards a logical analysis of the social contract*, Journal of Applied Logic **3** (2005), pp. 396–420.

Allowed, or enabled, that is the question

Giovanni Sileno [1]

Informatics Institute, University of Amsterdam
Science Park 900, 1098 XH Amsterdam, the Netherlands

Matteo Pascucci

Institute of Philosophy of the Slovak Academy of Sciences, v.v.i.
Klemensova 19, 811 09 Bratislava, Slovakia

Réka Markovich

Department of Computer Science, University of Luxembourg
2 Avenue de l'Université, 4365 Esch-Belval Esch-sur-Alzette, Luxembourg

Abstract

The formal analysis of normative systems has traditionally focused on their deontic dimension rather than on their potestative dimension; yet, a growing amount of works aims at shedding light on the notion of power, its norm changing potential and its general interactions with deontic concepts. The present article contributes to this line of inquiry by adopting the following perspective: a normative system can be metaphorically seen as an agent that allocates abilities (powers) in order to promote the fulfillment of certain desires (deontic directives), and in doing so regulates its behavioural domain. Our analysis emphasizes the instrumental nature of power, while clarifying the distinction between 'being allowed' and 'being enabled' and unveiling new patterns of interaction between deontic and potestative concepts. Operationally, we formulate this framework in terms of conditional rules, and provide a corresponding logic programming (ASP) implementation.

Keywords: Normative Systems; Permission; Power; Instrumental Reasoning; Answer Set Programming

1 Introduction

The idea of analysing normative systems in terms of deontic concepts has traditionally inspired many logic frameworks [1,12] and tends to overlook the potestative dimension present in normative discourse. This tendency can be also observed in computational systems that incorporate normative expressions, since they generally rely on the idea that (not) being granted *permission* to do

[1] Corresponding author: `g.sileno@uva.nl`.

a certain action upon the system is the same as (not) holding *power* to do it on the system. [2]

Yet, exceptions to this trend in the formal literature exist and are increasing—mainly in the tradition called the theory of normative positions [27] which is based on the characterisation of Hohfeld [13]. For instance, Lindahl [18] interprets power as a *possibility*, which can be either a permission, or a practical possibility, or a legal possibility. Jones and Sergot [14] treat power as a *count-as conditional*: within a given institution, some behaviours by certain parties count as ways of establishing normatively relevant states-of-affairs (e.g. within a department, secretaries' signatures count as their employers' signatures). Markovich [21] provides a definition of power as a *potential* involving an operator for *legal necessity*, which indicates that a party p has power on a party q when a certain behaviour of p brings legal consequences on some normative relation involving q and other parties. Dong and Roy [6,7] emphasize the relation-changing nature of power by defining it in a framework of *dynamic epistemic logic*, where actions available to some agents may affect the normative relations among others. Sileno and Pascucci [30] provide a definition of power in terms of *ability* and, in subsequent work [24], they build diagrams of opposition for various concepts of power (change-centered, outcome-centered and force-centered) and analyse their interactions. Kulicki, Trypuz and Sergot [15] use *labelled transition systems* in order to represent an agent's power to exercise a right in situations where such a right conflicts with those of other parties (e.g. a woman must be able to exercise her right to abortion despite doctors' conscience clause). Similarly, in the technical literature, new policy specification languages have been recently proposed, which includes potestative and deontic concepts, as for instance Symboleo [28] and eFLINT [31]. [3]

All mentioned approaches focus on representing the *way in which power produces changes in normative relations*; yet, they do not address the problem of characterizing the *reasons why power arises in institutional settings.* In other words, we lack a formal theory concerning the "instrumental" nature of power with respect to maintaining normative systems or producing a desired change in them. This is the problem we address in the present article and which, in turn, is related to the problem of appropriately characterizing the difference between the notions of permission ('being allowed') and power ('being enabled').

While permission and power indeed frequently come together, there are institutional settings in which they are activated by different conditions—proving the need for their ontological separation— as observed e.g. by Makinson [19], or Jones and Sergot [14]. For instance, according to canon law, people who are ordained priests retain the sacramental powers even when they leave priesthood: they are merely not allowed to exercise them. This shows that an ac-

[2] See e.g. `GRANT PRIVILEGE` to database users in MySQL, `Permit` effects in XACML rules for access control, `allow` directives in `.htaccess` for Apache webservers.

[3] Insights on how to structure a theory of power independently from a theory of deontic directives in computational settings can be found e.g. in Sileno, Boer and van Engers [29].

tion may produce (normative) effects even in presence of a prohibition that is meant to signal the undesirability of those effects. A similar observation can be drawn also on scenarios involving non-institutional (e.g. physical) actions that are normatively regulated while, obviously, no institutional power is associated to their performance. For instance, the action of smoking is in some contexts permitted although its performance does not involve any previous assignment of power. Another example of the separation between power and permission can be taken from markets operating on digital infrastructures, where enforcement occurs primarily *ex-ante*, in the tradition of authorization systems: transactions are allowed/enabled or not depending on certain conditions. However, markets also open the possibility of fraudulent schemes (i.e. complex behavioural patterns disrupting the normal functioning of the market), whose acknowledgement occurs mostly *ex-post.* Because of the impossibility of strict control (e.g. part of the scheme occurs off-chain), there may be outcomes which are prohibited, and yet possible, as they are in practice enabled by the infrastructure.

In order to analyse how power arises and to which extent the notions of power and permission diverge in normative settings, we introduce a formal framework where a normative system is seen as a collective agent, embodying the institution, that allocates abilities (corresponding to institutional powers) in order to promote the fulfillment of certain desires (corresponding to deontic directives). [4] This simplifying conceptual step is meant to ease the usage of practical reasoning constructs, generally discussed from an individual agent's stanpoint, rather than positing a perfect alignment between intentional and normative categories.

Our framework presents several mechanisms by means of which power originates; these mechanisms are expressed in the form of *conditional rules* and can be grouped into theories. The justification of the mechanisms comes from *reasonability* or *rationality* aspects of the norm-making process, which emerged already in the views of Georg Henrik von Wright [32,33] in terms of the suggested reading of deontic logic: a rational legislator does not create norms saying both $O\phi$ and $O\neg\phi$. Principles of rationality are also at the base of instrumental (or means-end) reasoning, as well as of engineering efforts relying on this, exemplified in control theory and goal-driven agents in AI, upon which we will build for this paper. The proposed formal theories will then have two readings: normative in the sense of the requirements rationality puts on the collective agent, and descriptive as specifying what these "rational" patterns are, implemented accordingly.

Moreover, specifications of power in legal settings generally consists of three dimensions (see e.g. Hart [11, p. 28]): *qualification* (requirements to be ascribed to a role), *performance* (manner and form in which the power is exercised), and *subject-matter* (variety of rights and duties which may be created or modified).

[4] The metaphor of institutions as agents echoes Hobbes' idea of the Leviathan and can be found also in recent technical works, such as Boella and van der Torre [3].

Since we map institutional powers to the abilities of a collective agent to *cause* changes in the normative system (and in particular, changes concerning potential changes), this paper will elaborate on what supports the creation or modification mechanisms associated to performances.

The structure of the article is as follows. Section 2 presents the formal framework, whose main ingredients are: the collective agent's desires, conditions holding in the world and causal connections between events and conditions. Section 3 provides conditional rules that can be used to build a theory for instrumental reasoning and that are grouped into patterns, according to the general mechanisms they represent. Section 4 provides additional patterns of rules corresponding to more specific institutional mechanisms. Finally, Section 5 presents an *answer set programming* (ASP) implementation of the framework.

2 Formal framework

We introduce a formal language $\mathcal{L}$ whose vocabulary and formulas are based on the syntax of the expressions used in ASP. The advantage of this choice is that the framework can be directly encoded into a program, as shown in the final part of the article.

2.1 Vocabulary

Each item in the vocabulary of $\mathcal{L}$ is associated with a *type* clarifying its meaning. Types can be *atomic* or *complex.* A complex type is either a *functional* type or a *union* type. A functional type $\mathtt{t}$ is denoted as $(\mathtt{t}_1, \mathtt{t}_2)$, where $\mathtt{t}_1$ and $\mathtt{t}_2$ are (possibly complex) types which respectively constitute the *input* and the *output* of $\mathtt{t}$. A union type $\mathtt{t}$ is denoted as $\mathtt{t}_1|\mathtt{t}_2$ and can be associated with symbols which are either of type $\mathtt{t}_1$ or of type $\mathtt{t}_2$. The atomic type boolean denotes the Boolean truth-values 0 and 1.

Object variables

We use upper case Latin letters for object variables. These variables are associated with two atomic types: condition (the type of *conditions* holding in the world, which convey factual information) and event (the type of events driven by agents, i.e. actions performed by them). Variables of each type are characterized by a particular notation in the following presentation: Cstand for conditions; A, B stand for events (sometimes with different subscripts making reference to names of individual agents that are part of the normative system, as in A_x, in which case they are actions driven by the individual agent indicated). We take object to be the type of all object variables, i.e. the union type event | condition. Object variables are denoted as X.

Object constants

We use lower case Latin letters for object constants. They are associated with the two types also used for object variables, namely condition and event. Notation is used accordingly, namely c for conditions and a for events (sometimes with different subscripts making reference to names of individual agents that are part of the normative system, as in a_x, in which case they are actions driven by the indicated agent).

Connectives

We use unary connectives for classical negation ($\neg$) and default negation ($\mathtt{not}$), binary connectives for classical conjunction ($\wedge$) and the ASP conditional ($\rightarrow$), and the universal quantifier ($\forall$). Unary connectives are associated with the type $(\mathsf{boolean}, \mathsf{boolean})$, binary connectives with the type $((\mathsf{boolean}, \mathsf{boolean}), \mathsf{boolean})$ and, for any object variable X, the expression $\forall X$ is associated with the type $(\mathsf{boolean}, \mathsf{boolean})$.[5] In the present context, the interpretation of $\rightarrow$ can be either *descriptive*, if one wants to characterize how an idealized normative system works, or *prescriptive*, if one wants to characterize how a normative system should be designed.

Function symbols

We use two binary function symbols $causes^+$ and $causes^-$ which take as input an event (first argument) and a condition (second argument). Their outputs are conditions. Thus, their type is: $((\mathsf{event}, \mathsf{condition}), \mathsf{condition})$. The function $causes^+$ represents a positive form of causation: an expression of the form $causes^+(A, C)$ reifies a causal mechanism binding action type A and condition C. More precisely, it indicates a condition according to which, by performing an action of type A, the agent triggers the consequent realization of C. The function $causes^-$ represents a negative form of causation: the expression $causes^-(A, C)$ reifies an inhibiting causal mechanism binding action type A and condition C. It indicates a condition according to which, by performing an action of type A, the agent inhibits the consequent realization of C (by any other means). We also use a unary function symbol *neg* which takes a condition as input and gives a condition (incompatible with C) as output.[6] Hence, its type is: $(\mathsf{condition}, \mathsf{condition})$.

Predicate symbols

Our framework involves reference to desires with a positive or negative attitude. We introduce two predicates Des^+ and Des^- which take a condition or an event as input and give a truth-value as output. Hence, their type is: $(\mathsf{object}, \mathsf{boolean})$. $Des^+(C)$ means that the collective agent has a positive attitude towards condition C (e.g. it prefers C to hold), whereas $Des^-(C)$ means that the collective agent has a negative attitude towards condition C (e.g. it prefers C to not hold). Analogous readings hold for $Des^+(A)$ and $Des^-(A)$. Moreover, we use a unary predicate *Holds* taking a condition as input and giving a truth-value as output. Hence, its type is: $(\mathsf{condition}, \mathsf{boolean})$. The meaning of an expression of the form $Holds(C)$ is that condition C holds.

[5] We stress that the connective $\rightarrow$ denotes a conditional operator typically used in ASP programs and behaving differently from material implication; for details, see [17]. Material implication (as well as the other classical connectives and $\exists$) is definable in $\mathcal{L}$ via the primitive connectives.

[6] The derivation mechanism (e.g. in our ASP implementation) may rely on intensional predicate functions, and therefore may not require to determine this incompatible condition.

Finally, we use a symbol for identity among objects (=). The type of = is $((\texttt{object}, \texttt{object}), \texttt{boolean})$.

2.2 Terms, formulas and instrumental theories

Definition 2.1 *Terms.* The set of terms (T_1, T_2, etc.) of $\mathcal{L}$ is the smallest set satisfying the following properties:

- every object variable and object constant is a term of $\mathcal{L}$;
- if T is a term whose type is $\mathsf{condition}$, then $neg(T)$ is a term of $\mathcal{L}$;[7]
- for every term T_1 of type event and term T_2 of type $\mathsf{condition}$, $causes^+(T_1, T_2)$ and $causes^-(T_1, T_2)$ are terms of $\mathcal{L}$.

Definition 2.2 *Formulas.* The set of atomic formulas of $\mathcal{L}$ is the smallest set satisfying the following properties:

- for every terms T_1 and T_2, $T_1 = T_2$ is a formula of $\mathcal{L}$;
- for every term T, $Des^+(T)$ and $Des^-(T)$ are formulas of $\mathcal{L}$;
- for every term T whose type is $\mathsf{condition}$, $Holds(T)$ is a formula of $\mathcal{L}$;
- if ϕ is a formula of $\mathcal{L}$, then so are $\neg\phi$ and $\texttt{not}(\phi)$;
- if ϕ and ψ are formulas of $\mathcal{L}$, then so are $\phi \wedge \psi$ and $\phi \rightarrow \psi$;
- if ϕ is a formula of $\mathcal{L}$, then so is $\forall X : \phi$, for X an object variable.

A formula is atomic iff it has one of the forms $T_1 = T_2$, $Des^+(T)$, $Des^-(T)$ or $Holds(T)$. We use $T_1 \neq T_2$ as an abbreviation for $\neg(T_1 = T_2)$. In the construction of terms and formulas the auxiliary symbols ':', '(' and ')' can be omitted according to binding conventions in the ASP syntax [17].

Definition 2.3 *Instrumental theories.* An instrumental theory Θ is a non-empty set of formulas which are either atomic or of the form $\phi \rightarrow \psi$ and, in the latter case, ψ is either of the form $Des^+(T)$ or of the form $Des^-(T)$. Within a theory, a formula of the form $\phi \rightarrow \psi$ is said to be a *conditional rule* and ψ is said to be the *target desire* of that rule. If Θ consists only of atomic formulas, then it is said to be an *atom-based* theory. Finally, a theory Θ_1 is an *expansion* of a theory Θ_2 iff $\Theta_1 \supseteq \Theta_2$.

Let Θ be an atom-based theory, namely a set of formulas describing either the collective agent's desires, or conditions holding in the world, or causal connections between events and conditions, or identity of objects. It is possible to expand Θ to a theory Θ' by adding conditional rules representing relevant mechanisms for instrumental reasoning that can be grouped into certain patterns. The next section will explain how this expansion can be performed. In all rules mentioned therein, free occurrences of variables should be understood as being in the scope of a universal quantifier, as in ASP rules [17].

[7] A technical remark: in the ASP implementation of the framework (Section 5) will only make use of terms where *neg* occurs at most once, since this is enough to encode the rules discussed in the article.

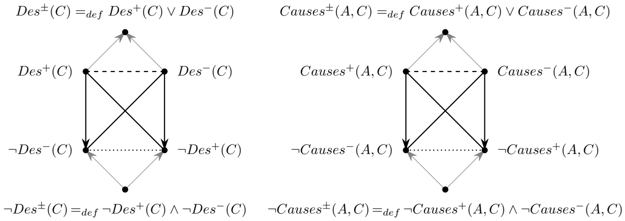

$Des^{\pm}(C) =_{def} Des^{+}(C) \vee Des^{-}(C)$
$Des^{+}(C)$
$Des^{-}(C)$
$\neg Des^{-}(C)$
$\neg Des^{+}(C)$
$\neg Des^{\pm}(C) =_{def} \neg Des^{+}(C) \wedge \neg Des^{-}(C)$
$Causes^{\pm}(A,C) =_{def} Causes^{+}(A,C) \vee Causes^{-}(A,C)$
$Causes^{+}(A,C)$
$Causes^{-}(A,C)$
$\neg Causes^{-}(A,C)$
$\neg Causes^{+}(A,C)$
$\neg Causes^{\pm}(A,C) =_{def} \neg Causes^{+}(A,C) \wedge \neg Causes^{-}(A,C)$

Fig. 1. Relations between elements of the language illustrated on hexagons of opposition. Labels of the hexagon on the right abuse the notation not to overload the image: e.g. $Causes^{+}(A,C)$ has to be read as $Holds(causes^{+}(A,C))$. Given two vertices v and u, an arrow from v to u indicates that u is a subalternate of v; a full line between v and u that v and u are contradictories; a dashed line between v and u that v and u are contraries; a dotted line between v and u that v and u are sub-contraries. The contrariety relation on the left hexagon holds only for collective agents corresponding to idealized normative systems; yet, we will also discuss scenarios in which a collective agent happens to have conflicting desires. The logic assumed for desires and causation is very simple and sufficient to serve the purposes of the article.

3 Patterns of conditional rules and theory expansion

3.1 Fundamental patterns

Suppose that the collective agent has certain desires with either positive or negative attitude with respect to a certain outcome (represented by condition C), and suppose that this condition currently is not in place. Moreover, suppose that action A is a causal mechanism producing C or inhibiting its production. Thus, we will start with an atom-based theory Θ that will contain one among $Des^{+}(C)$ and $Des^{-}(C)$, as well as one among $Holds(causes^{+}(A,C))$ and $Holds(causes^{-}(A,C))$, as well as $\neg Holds(C)$. The following patterns show how the collective agent's additional desires are triggered by the various combinations of these options. We stress an important point on the interpretation of conditional rules mentioned in the article. Under the descriptive reading of operator $\rightarrow$, a rule states that the target desire arises for the collective agent as an indication of a *sufficient* (rather than a *necessary*) instrument to get the desired outcome. Similarly, under the prescriptive reading of operator $\rightarrow$, a rule states that a *prima facie* (rather than an *all-things-considered*) duty arises for the collective agent with respect to the target desire. The reason behind this is that a sufficient instrument (a *prima facie* duty) does not have to be necessarily used (fulfilled) to achieve the intended goal, given that sometimes such an instrument (duty) might lead to unwelcome outcomes too. [8]

[8] We thank one of the reviewers for DEON for inquiring about this aspect of our framework.

(i) Means-end derivation

If you desire some (not present) condition to hold, and you have the ability to make this happen, you [should] desire to use such ability: [9]

$$\textbf{[1]} \quad Des^{+}(C) \land \neg Holds(C) \land Holds(causes^{+}(A, C)) \rightarrow Des^{+}(A)$$

The formula in the antecedent of the conditional is a conjunction composed of three elements that represent a common template in control theory and classic AI [10]: the *reference* (i.e. $Des^{+}(C)$) and the *current state* (i.e. $\neg Holds(C)$) define the trajectory, the *causal mechanism* (i.e. $Holds(causes^{+}(A, C))$) identifies the control that produces that trajectory. The consequent of the conditional indicates that the causal mechanism needs to be triggered. [11] Following a similar rationale, three additional rules can be identified:

$$\textbf{[2]} \quad Des^{-}(C) \land \neg Holds(C) \land Holds(causes^{+}(A, C)) \rightarrow Des^{-}(A)$$
$$\textbf{[3]} \quad Des^{-}(C) \land \neg Holds(C) \land Holds(causes^{-}(A, C)) \rightarrow Des^{+}(A)$$
$$\textbf{[4]} \quad Des^{+}(C) \land \neg Holds(C) \land Holds(causes^{-}(A, C)) \rightarrow Des^{-}(A)$$

Rule **2** is about the derivation of negative desires: actions which may trigger undesired events are undesired too. Rule **3** and Rule **4** deal with inhibiting mechanisms ($causes^{-}$), which are desired if they disable the occurrence of negatively desired events, undesired otherwise.

Next, we consider rules to represent situations in which the collective agent desires an instrument to obtain a certain goal.

(ii) Desire of instrument

If you desire some (not present) condition to hold, but you do not have the ability to make it happen, you [should] desire to create this ability. From a conceptual point of view, this rule works at a meta-level with respect to the previous ones; it indicates that before having the possibility of using a means to reach an end, we should have some means available. Similarly, *if you desire some (not present) condition to hold, and you have the ability to make it happen, you [should] desire not to remove this ability.* This second rule indicates

[9] The *should* in the various patterns indicates the normative (as opposed to descriptive) possible readings.

[10] See e.g. the general template of negative feedback [8, p. 8], the general architecture of goal-based agents [25, p. 56], or agent-programming language frameworks [4].

[11] Computational implementations based on practical reasoning principles typically set an *intentional bottleneck* constraint, because the agent triggers only one action to reach the desired outcome (e.g. the best one, in terms of costs and certainty). This passage, from possibly conflicting volitional elements (desires) to non-conflicting deliberative elements (intentions), is well-known in *beliefs-desires-intentions* (BDI) frameworks [5], and has strong connections to argumentative patterns (*necessary* vs *sufficient means*) [10,34,16], and to the distinction between *prima-facie* and all-things-considered obligations [2], of which we talked at the beginning of this section with respect to the two alternative readings. For simplicity, we keep these aspects out of our current scopes, although we acknowledge their relevance and plan to approach them in future works.

instead that when such a means is available, we should protect it. However, these patterns do not say anything about how the relevant causal mechanisms are created, and unnecessarily complicates the definition of $\mathcal{L}$. For this reason, we will rather focus on the following *realization* pattern in order to expand the theory Θ with mechanisms to create instruments for desired goals.

(iii) Creation of instrument

If you desire some (not present) condition to hold, and you do not have the ability to make it happen, but you have the ability to create such an ability, you [should] desire to use this ability.

$$\begin{aligned} Des^{+}(C) \wedge \neg Holds(C) \wedge \texttt{not}\ \exists A : Holds(causes^{+}(A, C)) \wedge \\ Holds(causes^{+}(B, causes^{+}(A, C))) \rightarrow Des^{+}(B) \end{aligned}$$

This pattern looks at the existing causal mechanisms and picks some action available to produce the target causal mechanism (enabling, or disabling a change).[12] This schema could in principle be applied recursively to take into account higher-order levels (e.g. causal mechanisms that create causal mechanisms that create...), but this is out of the scope for our present purposes.

The rules resulting from all combinations of desires, conditions and causal relations are the following:

(for 5–8) *let* ϕ *be* $\neg Holds(C) \wedge \texttt{not}\ \exists A : Holds(causes^{+}(A, C))$, *then*

[5] $\phi \wedge Des^{+}(C) \wedge Holds(causes^{+}(B, causes^{+}(A, C)))) \rightarrow Des^{+}(B)$

[6] $\phi \wedge Des^{+}(C) \wedge Holds(causes^{-}(B, causes^{+}(A, C)))) \rightarrow Des^{-}(B)$

[7] $\phi \wedge Des^{-}(C) \wedge Holds(causes^{+}(B, causes^{+}(A, C)))) \rightarrow Des^{-}(B)$

[8] $\phi \wedge Des^{-}(C) \wedge Holds(causes^{-}(B, causes^{+}(A, C)))) \rightarrow Des^{+}(B)$

(for 9–12) *let* ψ *be* $\neg Holds(C) \wedge \texttt{not}\ \exists A : Holds(causes^{-}(A, C))$, *then*

[9] $\psi \wedge Des^{+}(C) \wedge Holds(causes^{+}(B, causes^{-}(A, C)))) \rightarrow Des^{-}(B)$

[10] $\psi \wedge Des^{+}(C) \wedge Holds(causes^{-}(B, causes^{-}(A, C)))) \rightarrow Des^{+}(B)$

[11] $\psi \wedge Des^{-}(C) \wedge Holds(causes^{+}(B, causes^{-}(A, C))) \rightarrow Des^{+}(B)$

[12] $\psi \wedge Des^{-}(C) \wedge Holds(causes^{-}(B, causes^{-}(A, C))) \rightarrow Des^{-}(B)$

Rules **5-8** deal with the absence of causal mechanisms bringing about change; Rules **9-12** deal with the absence of inhibiting mechanisms.

(iv) Protection of instrument

We apply the same principle with the second pattern in (ii), concerning the protection of an existing relevant ability. We obtain therefore eight additional

[12] Note the use of connective `not` in this formula, which indicates that no such instrument has been found.

rules:

(for 13–16) *let* ϕ *be* $\neg Holds(C) \wedge Holds(causes^{+}(A, C))$, *then*

$$\textbf{[13]}\quad \phi \wedge Des^{+}(C) \wedge Holds(causes^{+}(B, neg(causes^{+}(A, C)))) \rightarrow Des^{-}(B)$$

$$\textbf{[14]}\quad \phi \wedge Des^{+}(C) \wedge Holds(causes^{-}(B, neg(causes^{+}(A, C)))) \rightarrow Des^{+}(B)$$

$$\textbf{[15]}\quad \phi \wedge Des^{-}(C) \wedge Holds(causes^{+}(B, neg(causes^{+}(A, C)))) \rightarrow Des^{+}(B)$$

$$\textbf{[16]}\quad \phi \wedge Des^{-}(C) \wedge Holds(causes^{-}(B, neg(causes^{+}(A, C)))) \rightarrow Des^{-}(B)$$

(for 17–20) *let* ψ *be* $\neg Holds(C) \wedge Holds(causes^{-}(A, C))$, *then*

$$\textbf{[17]}\quad \psi \wedge Des^{+}(C) \wedge Holds(causes^{+}(B, neg(causes^{-}(A, C))) \rightarrow Des^{+}(B)$$

$$\textbf{[18]}\quad \psi \wedge Des^{+}(C) \wedge Holds(causes^{-}(B, neg(causes^{-}(A, C)))) \rightarrow Des^{-}(B)$$

$$\textbf{[19]}\quad \psi \wedge Des^{-}(C) \wedge Holds(causes^{+}(B, neg(causes^{-}(A, C)))) \rightarrow Des^{-}(B)$$

$$\textbf{[20]}\quad \psi \wedge Des^{-}(C) \wedge Holds(causes^{-}(B, neg(causes^{-}(A, C)))) \rightarrow Des^{+}(B)$$

Patterns similar to (i), (iii) and (iv) can be constructed also *in presence* of the target condition C i.e. when the antecedent of a conditional rule includes $Holds(C)$. By making this amendment, one obtains twenty additional rules **[21–40]** (more precisely, for $\mathbf{1 \leq i \leq 20}$, rule $\mathbf{20 + i}$ is obtained by performing the mentioned amendment on rule **i**). These new rules are here not explicitly stated for reasons of space, but are present in the code. For more details, see Table 1.

3.2 Relaxations

All the patterns identified above require a desire to be present, either positively, either negatively, in order to trigger the derivation of new desires. Even if at the moment the agent does not desire something (the negation here stands as absence of desire), it is not the case that the agent has a negative volitional attitude towards that thing. This is a more relaxed condition, as indicated by the hexagon of opposition for desires (Fig. 1): $\neg Des^{-}(C)$ subalternates $Des^{+}(C)$. This relation identifies $\neg Des^{-}(C)$ as a *a necessary condition for* $Des^{+}(C)$ *to hold.* We may therefore explore the possibility of expanding theory Θ with new patterns involving the relaxed condition $\neg Des^{-}(C)$. We will distinguish two families of scenarios: contextual and direct anticipatory interventions.

3.2.1 Contextual interventions: preparing relevant abilities

In the first family of scenarios, the *anticipation* approach can be read as such: if the agent does not have a negative volitional attitude towards C, it is expected that at a certain point $Des^{+}(C)$ may hold, and so better be prepared by settling what will then be needed to bring about C. In other words, to motivate such endeavour it is sufficient to be committed to $\neg Des^{-}(C)$. In this way, once the desire comes to existence, relevant abilities are already in place for the agent. For instance, relaxing pattern **[5]** we obtain:

$$\textbf{[5*]}\quad \neg Des^{-}(C) \wedge \neg Holds(C) \wedge \texttt{not}\ \exists A : Holds(causes^{+}(A, C)) \wedge \\ Holds(causes^{+}(B, causes^{+}(A, C))) \rightarrow Des^{+}(B)$$

Following the same idea we can rewrite all patterns [**5**–**20**], [**25**–**40**] into relaxed forms [**5***–**20***], [**25***–**40***], which are about enabling or disabling (positive or negative) causal mechanisms that will be relevant when C—acknowledged as *desirable* (positively or negatively)—is eventually instantiated.

3.2.2 Direct interventions

The second family of scenarios concerns the core of the means-end derivation. Relaxing as before the premises concerning desire in e.g. Rule [**1**] and Rule [**2**], we obtain:

$$\textbf{[1*]}\quad \neg Des^{-}(C) \wedge \neg Holds(C) \wedge Holds(causes^{+}(A,C)) \rightarrow Des^{+}(A)$$
$$\textbf{[2*]}\quad \neg Des^{+}(C) \wedge \neg Holds(C) \wedge Holds(causes^{+}(A,C)) \rightarrow Des^{-}(A)$$

At further inspection, however, we observe that patterns concerning action avoidance (as [**2***]) have different practical implications than patterns concerning performance of actions ([**1***]). For instance, if C is recognized as *undesirable* (but possibly not as *undesired*), it is already acceptable that the agent should avoid performing an action A that would bring about C.[13] In contrast, if C is recognized as *desirable* (but possibly not as *desired*), it is less sound that the agent should perform an action to already bring about the condition C, just as we were in the positive desire case. A possible explanation for this intuition is that, independently of its effects in the world, executing an action carries always costs for the agent, whereas action avoidance generally plays a role only in plan selection, not in execution. In other words, the preference about applying the proposed relaxation between the two patterns seems to emerge out of principles of economy, entailing that instrumental reasoning common-sensically carries along a concurrent desire to select efficient plans (all other things being the same). This trail of thoughts is confirmed by observing that the avoidance case becomes unsound just as the performance case when there are few or no other plans available (or those that are available have much higher cost than the avoided plan).

Focusing for simplicity only on the qualitative dimension of these conditions, we can capture them as absence or presence of a *substitute* (equivalent and alternative) ability:

$$\begin{aligned} Holds(onecauses^{+}(A,C)) &=_{def} Holds(causes^{+}(A,C)) \wedge \\ &\quad \texttt{not}\ \exists B : Holds(causes^{+}(B,C)) \wedge A \neq B \\ Holds(manycauses^{+}(A,C)) &=_{def} Holds(causes^{+}(A,C)) \wedge \\ &\quad \exists B : Holds(causes^{+}(B,C)) \wedge A \neq B \end{aligned}$$

The patterns can then be rewritten as:

$$\textbf{[1**]}\quad \neg Des^{-}(C) \wedge \neg Holds(C) \wedge Holds(onecauses^{+}(A,C)) \rightarrow Des^{+}(A)$$
$$\textbf{[2**]}\quad \neg Des^{+}(C) \wedge \neg Holds(C) \wedge Holds(manycauses^{+}(A,C)) \rightarrow Des^{-}(A)$$

[13] Note that we are overlooking all problems related to defeasibility here. In the general case, the action A may still be required for satisfying concurrent desires of higher priority.

		Absence of target: $\neg Holds(C)$		Presence of target: $Holds(C)$	
	Available ability	$Des^{+}(C)$ or $\neg Des^{-}(C)$	$Des^{-}(C)$ or $\neg Des^{+}(C)$	$Des^{+}(C)$ or $\neg Des^{-}(C)$	$Des^{-}(C)$ or $\neg Des^{+}(C)$
`k.`	$causes^{-}(A, neg(C))$			$Des^{+}(A)$	$Des^{-}(A)$
`r.`	$causes^{+}(A, C)$	$Des^{+}(A)$	$Des^{-}(A)$		
`e.`	$causes^{+}(A, neg(C))$			$Des^{-}(A)$	$Des^{+}(A)$
`a.`	$causes^{-}(A, C)$	$Des^{-}(A)$	$Des^{+}(A)$		
`kk.`	$causes^{-}(B, neg(causes^{-}(A, neg(C))))$			$Des^{+}(B)$	$Des^{-}(B)$
`kr.`	$causes^{-}(B, neg(causes^{+}(A, C)))$	$Des^{+}(B)$	$Des^{-}(B)$		
`ke.`	$causes^{-}(B, neg(causes^{+}(A, neg(C))))$			$Des^{-}(B)$	$Des^{+}(B)$
`ka.`	$causes^{-}(B, neg(causes^{-}(A, C)))$	$Des^{-}(B)$	$Des^{+}(B)$		
`rk.`	$causes^{+}(B, causes^{-}(A, neg(C))$			$Des^{+}(B)$	$Des^{-}(B)$
`rr.`	$causes^{+}(B, causes^{+}(A, C))$	$Des^{+}(B)$	$Des^{-}(B)$		
`re.`	$causes^{+}(B, causes^{+}(A, neg(C))$			$Des^{-}(B)$	$Des^{+}(B)$
`ra.`	$causes^{+}(B, causes^{-}(A, C))$	$Des^{-}(B)$	$Des^{+}(B)$		
`ek.`	$causes^{+}(B, neg(causes^{-}(A, neg(C))))$			$Des^{-}(B)$	$Des^{+}(B)$
`er.`	$causes^{+}(B, neg(causes^{+}(A, C)))$	$Des^{-}(B)$	$Des^{+}(B)$		
`ee.`	$causes^{+}(B, neg(causes^{+}(A, neg(C))))$			$Des^{+}(B)$	$Des^{-}(B)$
`ea.`	$causes^{+}(B, neg(causes^{-}(A, C)))$	$Des^{+}(B)$	$Des^{-}(B)$		
`ak.`	$causes^{-}(B, causes^{-}(A, neg(C)))$			$Des^{-}(B)$	$Des^{+}(B)$
`ar.`	$causes^{-}(B, causes^{+}(A, C))$	$Des^{-}(B)$	$Des^{+}(B)$		
`ae.`	$causes^{-}(B, causes^{+}(A, neg(C)))$			$Des^{+}(B)$	$Des^{-}(B)$
`aa.`	$causes^{-}(B, causes^{-}(A, C))$	$Des^{+}(B)$	$Des^{-}(B)$		

Table 1

Overview of all instrumental patterns identified in section 2, in presence/absence of desired/undesired target, and considering presence/absence of four types of relevant abilities (`k`, `r`, `e`, `a`, standing respectively for *keep, reach, escape, avoid*). The relaxations from $Des^{+}(C)$ to $\neg Des^{-}(C)$ and from $Des^{-}(C)$ to $\neg Des^{+}(C)$ requires conditions on absence and on presence of substitute abilities (see section 2.2).

In words, performance is sustained by the absence of equivalent alternatives (although for an individual agent this desire is typically defeated due to economic reasons); action avoidance is sustained by the presence of equivalent alternatives (and generally it is not defeated as it does not incur in further costs). This idea will be applied on patterns **[1–4]** and **[21–24]**.

3.3 A framework of interventions

The 40 patterns can then be reorganized as follows. We first separate scenarios in which the target condition C is present from those in which it is absent, then we specify the attitude towards the target (positive or negative). The resulting organization is illustrated in Table 1. The evident symmetries suggest that further simplification of the notation is possible at a syntactic level, but this is beyond the scope of the present paper.

The practical derivation performed by the agent eventually depends on the available abilities, reified as causal connections, or possible interventions. For better readability, these abilities can labeled: `k` for *Keep* abilities, as e.g. $causes^{-}(A, neg(C))$, maintaining C in $Holds(C)$; `r` for *Reach* abilities, as e.g. $causes^{+}(A, C)$, producing C in $\neg Holds(C)$; `e` for *Escape* abilities, as e.g. $causes^{+}(A, neg(C))$, removing C in $Holds(C)$; `a` for *Avoid* abilities, as e.g.

$causes^{-}(A, C)$, inhibiting C in $\neg Holds(C)$.[14] This relabeling is useful here to denote in a more succint manner second-order abilities. For instance, `rr` will denote the ability to "reach" an ability to "reach" a certain condition—that is, $causes^{+}(B, causes^{+}(A, C))$.

4 Relevant institutional patterns

This section elaborates on how the machinery presented above can be applied to investigate patterns observable in institutional domains.

4.1 From (collective) agent to normative system

At this point, we want to interpret agentive attitudes in terms of a normative system, here taken as (i) a system of agents guided by (ii) a system of norms. Intuitively, the collective agent's desires would map to deontic directives, its abilities to potestative directives. Yet, two considerations are crucial in this passage. First, we need to take into account that there exist abilities which are *primitives* or given independently from the institution: either because they are physical abilities proper of individuals, or because they are (recognized) institutional abilities provided by some other institution. Second, actions A performed by the collective agent map to actions performed by individual agents *for the sake of the institution*. However, because they are autonomous, individuals may still perform actions for other purposes (e.g. for their own interests). We will utilize subscripts to distinguish individuals, e.g. A_x would be an event driven by agent x. We will now consider a few relevant patterns to show potential applications of the proposed framework.

4.2 Protected liberty

In the normative system literature, a distinction is usually made between permissions (and/or liberties) which are explicitly declared, and those which are derived from the absence of relevant obligations or prohibitions. Various authors[15] have argued that a *permissive norm* issuing a "strong permission" bring along additionally mechanisms, that can be overall reorganized as:

[14] This framework is similar to taxonomies presented in other disciplines. For instance, works in agent-based programming distinguish *maintenance*, *achievement*, *remedy*, and *avoidance* goals. Similarly, in psychology, Ogilvie and Rose [23] introduce the *prevent-acquire-cure-keep* (PACK) framework to classify explanations given by people about their own behaviour. The PACK framework of motives however takes also into account the positive or negative attitude of the agent towards the target. For instance *acquire* (A) is always about reaching a positive outcome; the agent is not deemed to reach (purposely) a negative outcome.

[15] For instance, Makinson [19] observes that explicit permission "appears to be needed by real-life normative systems that change over time, as a device for limiting the interpretation of obligations and preventing their proliferation." Together with Alchourrón and Bulygin, he sees this practice as "to limit the authority of subordinate instances to create new norms" that go against the given permission. Additional mechanisms are summarized by Sartor [26], building upon Hart (for the protection coming with explicit permissions), Alexy and Pettit (for protected freedoms), and Sen (for the creation of effective capability in presence of permission). More recently, Markovich and Roy provide a logical formalization of the freedom of thought pointing out all the protective layers [22].

- a *practical protection*: the prohibition of interference against performance addressing all other social participants;
- an *institutional protection*: the disability for subordinate regulators to produce directives conflicting with that norms;

A special case is then that of institutional actions; here explicit liberty (e.g. right to marry discussed by Markovich [20]) also implies an obligation of the normative system to follow along the consequence of the action, i.e.

- an *institutional instrument*, i.e. a power for the addressee of the liberty to require a certain performance from the normative system.

A non-intervention by the normative system (the only one that can keep track of effects of institutional actions as marrying) would count as an interference.

For instance, let us consider a directive as *x is free to marry, as well as free not to marry.* The practical protection function entails that someone should not prohibit or interfere with x's marrying, as well as nobody should oblige or control x in this sense. The institutional protection entails that subordinate regulators cannot change this directive. The institutional instrument entails that x should be enabled to marry if x wished to (the institution being the only agency able to produce this institutional outcome).

4.2.1 Application of the proposed framework

Rephrasing this discussion in agentive terms, a weak liberty would map to having $\neg Des^{\pm}$ derived from the absence of other volitional attitudes (i.e. issued by some form of default negation), whereas a strong liberty would map to stating the attitude $\neg Des^{\pm}$ (entailing strongly negated statements). The directive expressed in the example is an explicit expression of strong liberty, in the form of a $\neg Des^{\pm}(C)$ position:

$$\neg Des^{\pm}(C) \rightarrow \neg Des^{+}(C) \wedge \neg Des^{-}(C)$$

Looking at Table 1, this entails that both conclusions in the first and the second column are potentially relevant. The role of the constraints on relaxation (section 2.2) based on substitute abilities becomes here particularly relevant, as they prevent to conclude opposite desires.

As a first validation, let us check whether our framework captures what is expected from the legal literature. With respect to *practical protection*, we have:

- If an event A_y can inhibit the outcome C (e.g. y can interfere with x's marrying), knowing that A_x also can do it (x can refrain from marrying), entails that A_y is undesired (Table 1, `a`, first column).
- If an event A_y can bring about the outcome C (e.g. y has the ability to control marrying besides x), then A_y is undesired (Table 1, `r`, second column).

With respect to *institutional protection*:

- The institution needs to protect x's ability by disabling the possibility to remove it (Table 1, `kr`, first column).

Note that in this case there are no substitute abilities (only the institution can intervene on this matter) and this entails that the conclusion is a positive desire. With respect to *institutional instrument*,

- The institution needs to put in place mechanisms so that x's marrying is eventually acknowledged (Table 1, `rr`, first column)

Note that the framework concludes also that the institution should abstain to create or keep instruments that may interfere with the marriage (Table 1, `ka` and `ra`, first column), as long as x is expected to have the ability of abstaining from marrying. Dually, our framework would suggest that if marriages may be combined (and individuals have no ability to abstain from marrying), dedicated institutional instruments should be created to empower individuals to stop marriages to occur. The legal theoretical understanding of such a conclusion could be of course subject to a detailed discussion, but we only refer to the openness of this question here.

5 Implementation

We have implemented a version of the framework in *answer set programming* (ASP) [17].[16] We will report here excerpts of the code, the full version is publicly available.[17] The notation used in the code is slightly adapted from the formal framework, following syntactic conventions of the ASP syntax (`-` represents $\neg$, `posdes` represents Des^+, `poscauses` represents $causes^+$, `nodes` represents $\neg Des^\pm$, all predicates start with lower case letters, `:-` replaces $\rightarrow$ switching antecedent and consequent, etc.). For instance, rule [**1**] becomes:

```
posdes(A) :- posdes(C), -holds(C), holds(poscauses(A, C)). % 1 (r)
```

Stating that there is no available cause of a condition requires using *default negation* (`not`), as in the second line of the code below:

```
holds(some_poscauses(C)) :- holds(poscauses(_, C)).
posdes(B) :- posdes(C), -holds(C), not holds(some_poscauses(C)),
             holds(poscauses(B, poscauses(A, C))). % 5 (rr)
```

Negated conditions require a function `neg` operating at the level of terms, e.g.:

```
negdes(A) :- posdes(C), holds(C),
             holds(poscauses(A, neg(C))). % 21 (e)
negdes(B) :- posdes(C), -holds(C), holds(poscauses(A, C)),
             holds(poscauses(B, neg(poscauses(A, C)))). % 13 (er)
```

In order to take into account anticipatory patterns, following the relaxation discussed in Section 3, we first need to define relations corresponding to subalternation in the deontic hexagon of Figure 1:

```
-negdes(C) :- posdes(C). -posdes(C) :- negdes(C).
```

[16] ASP is a declarative programming paradigm based on a *stable-model* semantics [9], oriented towards NP-hard search problems, and increasingly used to model and solve problems in research and industry in a wide range of application domains.

[17] `https://github.com/gsileno/abilities-desires-asp`

```
-negdes(C) :- nodes(C). -posdes(C) :- nodes(C).
```

The two constraints we require for the relaxation are encoded as:

```
holds(one_poscauses(A, C)) :-
   holds(poscauses(A, C)), not holds(many_poscauses(A, C)).
holds(many_poscauses(A, C)) :-
   holds(poscauses(A, C)), holds(poscauses(B, C)), A != B.
```

We can then rewrite the patterns in the relaxed form:

```
posdes(A) :- -negdes(C), -holds(C),
   holds(one_poscauses(A, C)). % 1** (r)
negdes(A) :- -posdes(C), -holds(C),
   holds(many_poscauses(A, C)). % 2** (r)
```

For the combinatorial exploration, we specify that all conditions present in the program may hold or not:

```
{holds(C)} :- condition(C).
```

We also introduce four integrity constraints, one for each type of causal connection. For instance, the potential transition reified in $causes^{+}(A, C)$ requires C not to hold (otherwise there would be no change):

```
-holds(C) :- holds(poscauses(A, C)), condition(C).
```

To reason about the absence of conditions, we need to introduce a closed-world assumption relying on default negation:

```
-holds(C) :- not holds(C), condition(C).
```

With these rules, we can specify a certain normative configuration (deontic and potestative directives mapped to desires and abilities) and automatically derive what theoretically entailed by the instrumental reasoning patterns identified above. The result may confirm and possibly extend normative constructs discussed in the literature. For instance, supposing that x's marrying (event denoted as `m_x`) is permitted (if interpreted as facultativeness, in our framework it is encoded as `nodes(m_x)`), we can check the associated normative consequences in different potestative configurations. For instance, suppose that the authority a has the institutional power to create the power for x to marry, i.e.:

```
nodes(m_x). holds(poscauses(a_a, poscauses(a_x, m_x))).
```

Indeed, we derive that the authority should exercise its power, i.e. `posdes(a_a)`. Now let us assume that this power already exists:

```
nodes(m_x). holds(poscauses(a_x, m_x)). holds(poscauses(a_y, m_x)).
```

If x and y have both the ability to bring about x's marriage (for instance forcing x to do so), y is forbidden to do so, i.e. we derive `negdes(a_y)`. However, this scenario shows also a limitation of the current formalization: as it does not allow for distinguishing individuals driving actions (e.g. legitimate from illegitimate), it generates also a prohibition upon x. We leave this extension to future work.

6 Conclusion

The paper introduces a framework to investigate the relationships between potestative and deontic categories, looking at institutions as collective agents and exploiting instrumental reasoning patterns. The framework allows for performing a combinatorial exploration of several patterns of interactions between powers, obligations, prohibition and permissions. As an example, by performing the derivation on a strong permission/liberty, we entailed several mechanisms discussed (often separately, and by distinct authors) in the normative systems' literature. Yet, we acknowledge that these results are just initial with respect to the potential applications of the framework. At the moment, our analysis is a-temporal. We primarily identify desires "rationally" holding at a certain moment of time, including those concerning the modification of abilities, without being concerned of solving conflicts that may emerge. In future work, we aim to add to the present framework deliberative and causal/temporal modules.

The deliberative module will serve to select a set of (non-conflicting) intentions based upon the existing (possibly conflicting) desires. This problem has connections with the distinction between *prima-facie* vs actual obligations. Indeed, a normative system does not consists only of mechanisms for allocating powers, but also of substantial and procedural constraints on such allocation. In our current formalization, these constraints may be captured as directives which, in the moment of allocation/derivation, would determine conflicts. The causal/temporal module will enable reasoning about the effects of events (including performances driven by intention). Once these modules are integrated with the present framework, we could reason on the overall institutional dynamics. For instance, the need for institutional protection will be derived automatically after an institutional instrument has been decided and created.

Complementary to these extensions, further effort is needed to identify how to unveil institutionally relevant generic mechanisms, as for instance the power of declaring the occurrence of a violation, and of requiring interventions from an enforcer. More fundamentally, the prescriptive reading opens up also to the use of the framework for assistive technologies: given a certain normative system, how could this be improved?

Acknowledgments Giovanni Sileno was partly funded by the Dutch Research Council (NWO) for the HUMAINER AI project (KIVI.2019.006). Matteo Pascucci was supported by the Štefan Schwarz Fund (2020-2023) for the project "A fine-grained analysis of Hohfeldian concepts" and by VEGA grant n. 2/0125/22 "Responsibility and modal logic". Réka Markovich was supported by the Fonds National de la Recherche Luxembourg through the project INTEGRAUTO (INTER/AUDACE/21/16695098).

Authors' contribution The ideas at the basis of this work rely on extended discussion among the three authors. As far as the article production is concerned, Sections 3, 4, 5 and 6 are mainly due to Giovanni Sileno, Section 2 is mainly due to Matteo Pascucci and Section 1 is equally due to all the authors.

References

[1] Åqvist, L., *Deontic logic*, in: D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, 2002 pp. 147–264.

[2] Asher, N. and D. Bonevac, *"Prima facie" obligation*, Studia Logica **57**, pp. 19–45.

[3] Boella, G. and L. van der Torre, *Constitutive norms in the design of normative multiagent systems*, in: F. Toni and P. Torroni, editors, *Computational Logic in Multi-Agent Systems* (2006), pp. 303–319.

[4] Bordini, R., J. Hübner and M. Wooldridge, "Programming Multi-Agent Systems in AgentSpeak using Jason," Wiley Series in Agent Technology, Wiley, 2007.

[5] Bratman, M., "Intention, plans, and practical reason," CSLI publications, 1987.

[6] Dong, H. and O. Roy, *Dynamic logic of power and immunity*, in: J. Baltag, A. Seligman and T. Yamada, editors, *Logic, Rationality, and Interaction*, 2017 pp. 123–136.

[7] Dong, H. and O. Roy, *Dynamic logic of legal competences*, Journal of Logic, Language and Information **30** (2021), pp. 701–724.

[8] Doyle, J. C., B. A. Francis and A. R. Tannenbaum, "Feedback control theory," Macmillan Publishing Co., 1990.

[9] Gelfond, M. and V. Lifschitz, *The stable model semantics for logic programming*, Proceedings of International Logic Programming Conference and Symposium (1988), pp. 1070–1080.

[10] Hare, R. M., "Practical inferences," MacMillan Press Ltd, 1971.

[11] Hart, H. L. A., "The concept of law," Oxford University Press, 1994, second edition.

[12] Hilpinen, R. and P. McNamara, *Deontic logic: a historical survey and introduction*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, 2013 pp. 1–134.

[13] Hohfeld, W. N., *Fundamental legal conceptions as applied in judicial reasoning*, The Yale Law Journal **26** (1917), pp. 710–770.

[14] Jones, A. and M. Sergot, *A formal characterisation of institutionalised power*, Journal of IGPL **4** (1996), pp. 427–443.

[15] Kulicki, P., R. Trypuz and M. Sergot, *Who is obliged when many are involved? Labelled transition system modelling how obligation arises*, Artificial Intelligence and Law **29** (2021), pp. 395–415.

[16] Lewinski, M., *Practical argumentation as reasoned advocacy*, Informal Logic **37** (2017), pp. 85–113.

[17] Lifschitz, V., *What is answer set programming?*, Proceedings of the AAAI Conference on Artificial Intelligence (2008).

[18] Lindahl, L., "Position and Change: A Study in Law and Logic," Synthese Library, Springer, 1977.

[19] Makinson, D., *On the formal representation of rights relations*, Journal of Philosophical Logic **15** (1986), pp. 403–425.

[20] Markovich, R., *No match-making but biconditionals. agents and the role of the state in legal relations*, in: A. Rotolo, editor, *Legal Knowledge and Information Systems 2015 – Proceedings of the 29th JURIX Conference. Frontiers in Artificial Intelligence and Applications*, IOS Press, 2015 pp. 161–164.

[21] Markovich, R., *Understanding Hohfeld and formalizing legal rights: the Hohfeldian conceptions and their conditional consequences*, Studia Logica **108** (2020), pp. 129–158.

[22] Markovich, R. and O. Roy, *A logical analysis of freedom of thought*, in: P. P. Fenrong Liu, Alessandra Marra and F. V. D. Putte, editors, *Deontic Logic and Normative Systems DEON 2020/21 proceedings*, College Publications, 2021 pp. 245–260.

[23] Ogilvie, D. M. and K. M. Rose, *Self-with-other representations and a taxonomy of motives: two approaches to studying persons*, Journal of personality **63** (1995), pp. 643–79.

[24] Pascucci, M. and G. Sileno, *A formal, diagrammatic and operational study of normative relations*, Journal of Logic and Computation (2023), pp. 1–30.

[25] Russell, S. J. and P. Norvig, "Artificial Intelligence: a modern approach," Pearson, 2009, 3 edition.

[26] Sartor, G., *Fundamental legal concepts: A formal and teleological characterisation*, Artificial Intelligence and Law **14** (2006), pp. 101–142.

[27] Sergot, M., *Normative positions*, Handbook of deontic logic and normative systems **1** (2013), pp. 353–406.

[28] Sharifi, S., A. Parvizimosaed, D. Amyot, L. Logrippo and J. Mylopoulos, *Symboleo: Towards a Specification Language for Legal Contracts*, in: *IEEE International Requirements Engineering Conference (RE'20), RE@Next! track*, 2020.

[29] Sileno, G., A. Boer and T. van Engers, *Towards a computational theory of action, causation and power for normative reasoning*, Proceedings of the 32nd International Conference on Legal Knowledge and Information Systems (JURIX 2019) (2019).

[30] Sileno, G. and M. Pascucci, *Disentangling deontic positions and abilities: a modal analysis*, in: F. Calimeri, S. Perri and E. Zumpano, editors, *Proceedings of the 35th Edition of the Italian Conference on Computational Logic (CILC 2020)*, 2020, pp. 36–50.

[31] van Binsbergen, L. T., L.-C. Liu, R. van Doesburg and T. van Engers, *eFLINT: a Domain-Specific Language for Executable Norm Specifications*, in: *GPCE 2020: Proceedings of the 19th ACM SIGPLAN International Conference on Generative Programming*, 2020.

[32] von Wright, G. H., *Is and ought*, in: M. C. Doeser and J. N. Kraay, editors, *Facts and Values: Philosophical Reflections from Western and Non-Western Perspectives*, Springer, Dordrecht, 1986 p. 31–48.

[33] von Wright, G. H., *Is there a logic of norms?*, Ratio Juris **4** (1991), p. 265–283.

[34] Walton, D., *Evaluating practical reasoning*, Synthese **157** (2007), pp. 197–240.

Inconsistent precedents and deontic logic

Ilaria Canavotto[1]

Philosophy Department, University of Maryland, College Park
4300 Chapel Drive, College Park
MD 20742, USA

Abstract

Computational models of legal precedent-based reasoning developed in the field of Artificial Intelligence and Law are typically based on the simplifying assumption that the background set of precedent cases is consistent. Besides being unrealistic in the legal domain, this assumption is problematic for recent promising applications of these models to the development of explainable Artificial Intelligence methods. In this paper I explore a model of legal precedent-based reasoning that, unlike existing models, does not rely on the assumption that the background set of precedent cases is consistent. The model is a generalization of the reason model of precedential constraint. I first show that the model supports an interesting deontic logic, where consistent obligations can be derived from inconsistent case bases. I then provide an explanation of this surprising result by proposing a reformulation of the model in terms of cases that support a new potential decision and cases that conflict with it.

Keywords: Legal case-based reasoning, reason model, inconsistent case bases, explainable artificial intelligence.

1 Introduction

Suppose that it is Monday morning and we post a homework assignment for our logic class with due date on Friday at 4:00pm. Right after we post the assignment, a student, Ann, writes us an email asking for an extension because she is going to be at a conference for the entire week. Suppose that we grant her the extension. On Wednesday another student, Bob, writes us an email asking for an extension because he is going to be at a conference on Thursday and Friday. We decide not to grant him the extension. Bob writes us back complaining that Ann, who was in a similar situation, was granted an extension. Did we have an obligation to grant Bob an extension given how we decided Ann's case? Or were we permitted to decide Bob's case the way we did?

The question of how previous authoritative decisions, or precedent cases, constrain future decisions has been addressed especially in legal theory in the common law tradition. According to the common law doctrine of precedent,

[1] icanavot@umd.edu.

the decisions of earlier courts constrain the decisions of later courts through the requirement that later decisions ought to be consistent with precedent decisions. But what, exactly, does "consistency" mean? Besides being a traditional problem in legal theory, this has become, through the development of the reason model of constraint by Horty and Bench-Capon [11,14], a central concern in the field of Artificial Intelligence and Law (AI and Law) as well.

The reason model, which builds on Lamond's theory of precedential constraint [15], supplements a factor-based representation of legal cases in the style of early models of legal case-based reasoning like HYPO [3] and CATO [1] with a priority ordering between sets of factors representing the strength of the reasons underlying the decisions of different courts. With respect to earlier proposals based on similar ideas [5,17,19], the key innovation is that this priority ordering is used to define a notion of consistency, and so a notion of constraint.

This has led to a number of developments in AI and Law that aim at refining the analysis of constraint by tackling, for instance, factors that can have multiple values [12,17,21], framework precedents [20], or issues [6]. A problem that has not been taken up in this literature, however, is that the reason model notion of consistency presupposes that the background set of precedent cases is consistent to start with. Besides being unrealistic in the legal domain, this assumption is also problematic for recent promising applications of models of precedential constraint to the development of AI systems that can learn and reason about normative information in a way that is explainable. The key idea behind these approaches is to interpret training data sets as sets of precedent cases and then use the relevant model of precedential constraint to either build interpretable systems [7,8,9] or construct post hoc explanation algorithms for machine learning systems for binary classification [18]. One challenge in implementing this idea is that training data is typically inconsistent.

Horty [11] briefly mentions a generalization of the reason model notion of constraint that applies to inconsistent case bases as well. Yet, the idea is only presented and not explored in any detail. My aim in this paper is to take Horty's suggestion and study how, exactly, according to the generalized notion of constraint, inconsistent case bases constrain future decisions and generate permissions and obligations for future courts. [2]

I proceed as follows. In Section 2, I review some basic definitions and present the generalized reason model of constraint. In Section 3, I define what it means, in this framework, for it to follow from a possibly inconsistent case base that a decision is obligatory or permitted and show that the resulting notions of permission and obligation support a simple conflict-free deontic logic. This result is promising but surprising: how does the generalized reason model extract consistent requirements from an inconsistent set of precedent cases? In

[2] A different approach to the problems presented by inconsistent case bases can be found in recent work by Peters and colleagues [16] and by van Woerkom and colleagues [22], which develop Prakken's and Ratsma's proposal to use a version of the reason model to analyze how machine learning systems base their decisions on training data [18].

Section 4, I present a reformulation of the generalized reason model that will provide me with a way to answer this question in Section 5. Finally, Section 6 concludes.

2 The generalized reason model

The reason model represents cases as consisting of three elements: a fact situation presented to the court; an outcome, which can be either a decision for the plaintiff or a decision for the defendant; and a rule that justifies the outcome on the basis of a reason that holds in the considered situation. I start by reviewing the definitions of these elements.

A *fact situation* is a set of facts that are legally relevant, called *factors*. Factors are assumed to have polarities: every factor favors either the plaintiff, denoted with π, or the defendant, denoted with δ. We take $\mathcal{F}^{\pi} = \{f_1^{\pi}, \ldots, f_n^{\pi}\}$ to be the set of factors favoring the plaintiff, $\mathcal{F}^{\delta} = \{f_1^{\delta}, \ldots, f_m^{\delta}\}$ to be the set of factors favoring the defendant, and $\mathcal{F} = \mathcal{F}^{\pi} \cup \mathcal{F}^{\delta}$ to be the set of all factors. Where s is one of the two sides, we will use $\overline{s}$ to represent the other, so $\overline{s} = \pi$ if $s = \delta$ and $\overline{s} = \delta$ if $s = \pi$. Where X is a fact situation, $X^s = X \cap \mathcal{F}^s$ is the set of factors from X that favor the side s. For example, if $X_1 = \{f_1^{\pi}, f_2^{\pi}, f_1^{\delta}\}$, then $X_1^{\pi} = \{f_1^{\pi}, f_2^{\pi}\}$ and $X_1^{\delta} = \{f_1^{\delta}\}$.

Next, a *reason for the side* s is a non-empty set of factors uniformly favoring s; a *reason* is then a non-empty set of factors uniformly favoring *a* side. We say that a reason U *holds* in a fact situation X whenever $U \subseteq X$ and that U is *at least as strong as* another reason V favoring the same side as U whenever $V \subseteq U$. To illustrate, the sets $\{f_1^{\pi}\}$ and $\{f_1^{\pi}, f_2^{\pi}\}$ are reasons for π that hold in the previous fact situation X_1 and such that $\{f_1^{\pi}, f_2^{\pi}\}$ is at least as strong as $\{f_1^{\pi}\}$.

We can now define a *rule* as a statement of the form $U \rightarrow s$, where U is a reason for the side s. Intuitively, $U \rightarrow s$ represents a defeasible rule that, roughly, says that, if U holds in a fact situation, then the court has a *pro tanto* reason to decide that situation for s. For any rule $r = U \rightarrow s$, we let $\mathit{premise}(r) = U$ and $\mathit{conclusion}(r) = s$. We say that r is *applicable in a fact situation* X whenever its premise holds in X, that is $\mathit{premise}(r) \subseteq X$.

At this point, we can define a *case* as any triple of the form $\langle X, r, s \rangle$, where X is a fact situation, r is a rule applicable in X and whose conclusion is s, and s is either π or δ. For any case $c = \langle X, r, s \rangle$, we set $\mathit{facts}(c) = X$, $\mathit{rule}(c) = r$, and $\mathit{outcome}(c) = s$. Since the rule of a case justifies the outcome on the basis of the reason that forms its premise, in the following I will indifferently say that a case is decided on the basis of either its rule or the premise of its rule.

Finally, a *case base* Γ is simply a set of cases. A case base represents the set of precedent cases that constrain the decisions of future courts. How does it do this? Well, the reason model is based on two key ideas: first, that every case decided by a court induces a priority ordering among reasons and, second, that the decisions taken by later courts ought to be consistent with the priority ordering induced by precedent cases.

To make the previous ideas precise, we start by defining the priority ordering induced by a case:

Definition 2.1 [Priority ordering induced by a case.] Where $c = \langle X, r, s\rangle$ is a case, the priority ordering $<_c$ induced by c is defined by setting, for any pair of reasons $U \subseteq \mathcal{F}^{\overline{s}}$ and $V \subseteq \mathcal{F}^{s}$: $U <_c V$ if and only if $U \subseteq X$ and $\mathit{premise}(r) \subseteq V$.

To illustrate, let c_1 be the case $\langle X_1, r_1, \pi\rangle$, where X_1 is as above and $r_1 = \{f_1^\pi\} \rightarrow \pi$. The idea behind Definition 2.1 is that c_1 reveals that, according to the court, the reason $\{f_1^\pi\}$ has higher priority than every reason for δ that holds in X_1—i.e., $\{f_1^\delta\}$—and that every reason for π that is at least as strong as $\{f_1^\pi\}$, for instance $\{f_1^\pi, f_2^\pi, f_3^\pi\}$, also has higher priority than every such reason. It is worth noting that Definition 2.1 ensures that the ordering $<_c$ is asymmetric: there are no reasons U and V such that $U <_c V$ and $V <_c U$.

Having defined the notion of a priority ordering induced by a case, we can now lift it to a corresponding notion of a priority ordering induced by a case base Γ by simply requiring that a reason have higher priority than another according to Γ just in case Γ contains a case that induces that priority:

Definition 2.2 [Priority ordering induced by a case base.] Where Γ is a case base, the priority ordering $<_\Gamma$ induced by Γ is defined by setting, for any pair of reasons U and V: $U <_\Gamma V$ if and only if there is a case c in Γ such that $U <_c V$.

Observe that, unlike Definition 2.1, Definition 2.2 does not force $<_\Gamma$ to be asymmetric: there may be reasons U and V such that $U <_\Gamma V$ and $V <_\Gamma U$. This happens when some cases in Γ support conflicting information about the priority ordering among reasons. Such cases make Γ inconsistent. More precisely, let us define the notion of an inconsistency in Γ as follows:

Definition 2.3 [Inconsistency in Γ.] Where Γ is a case base, an inconsistency in Γ is any pair of reasons U and V such that $U <_\Gamma V$ and $V <_\Gamma U$.

We can then define the notions of an inconsistent and of a consistent case base in the expected way:

Definition 2.4 [Inconsistent and consistent case base.] A case base Γ is inconsistent when there is an inconsistency in Γ and consistent otherwise.

So, if c_1 is as before and c_2 is the case $\langle X_2, r_2, \delta\rangle$, where $X_2 = \{f_1^\pi, f_1^\delta, f_2^\delta\}$ and $r_2 = \{f_1^\delta\} \rightarrow \delta$, then the case base $\Gamma_1 = \{c_1, c_2\}$ is inconsistent: In fact, as we have seen above, according to the priority ordering derived from c_1, the reason $\{f_1^\pi\}$ has higher priority than the reason $\{f_1^\delta\}$, while, as the reader can easily verify, the opposite is true according to the priority ordering derived from c_2. Since c_1 and c_2 belong to Γ_1, the reasons $\{f_1^\pi\}$ and $\{f_1^\delta\}$ thus form an inconsistency in Γ_1: $\{f_1^\delta\} <_{\Gamma_1} \{f_1^\pi\}$ and $\{f_1^\pi\} <_{\Gamma_1} \{f_1^\delta\}$.

Now, in the previous example, the case base Γ_1 is inconsistent in a way that is so obvious that it would be striking if any court actually had to work with a case base like it. But, in real life, case bases are much more complex than Γ_1 and it is not at all unusual that some precedents pull in different directions. The question *How do inconsistent case bases constrain?* thus becomes pressing. The reason model notion of constraint does not allow us to pose—let alone answer—this question. To see this, recall that, according to the reason model, decisions of later courts ought to preserve consistency of the underlying case base. This idea is modeled in two steps. First, we characterize the rules that a court is permitted to use to justify a decision:

Definition 2.5 [Reason model: permitted rules] Let Γ be a consistent case base. Then, against the background of Γ, the court is permitted to decide the fact situation X for the side s on the basis of a rule r, applicable in X and favoring s, just in case the augmented case base $\Gamma \cup \{\langle X, r, s\rangle\}$ is consistent.

And then, given this notion of permission, we say that the reason model constrains the court to reach a decision on the basis of some applicable rule that is permitted according to the model. The problem with inconsistent case bases is that Definition 2.5 explicitly requires that the underlying case base be consistent. Even worse, simply dropping this requirement would not give us a sensible account of how inconsistent case bases constrain—given an inconsistent case base, there would be no permitted way at all to decide any new fact situation, which is absurd.

Fortunately, however, there is another way to generalize the reason model notion of constraint so that it applies to inconsistent case bases as well. The idea, which was suggested in [11, p.15], is that, rather than being required to preserve consistency of a consistent case base, courts should be required to introduce no new inconsistencies into a possibly inconsistent case base. What is a new inconsistency? Well, let Γ and Γ' be two case bases such that Γ' extends Γ. Then a new inconsistency in Γ' with respect to Γ is simply an inconsistency present in Γ' but not in Γ:

Definition 2.6 [New inconsistency with respect to Γ] Let Γ and Γ' be two case bases such that $\Gamma \subset \Gamma'$. Then a pair of reasons U and V is a new inconsistency in Γ' with respect to Γ if and only if it is the case that $U <_{\Gamma'} V$ and $V <_{\Gamma'} U$ but it is not the case that $U <_{\Gamma} V$ and $V <_{\Gamma} U$.

With the notion of a new inconsistency in place, we can, first, generalize the reason model notion of a permitted rule as follows:

Definition 2.7 [Generalized reason model: permitted rule.] Against the background of a case base Γ, the court is permitted to decide the fact situation X for the side s on the basis of the rule r, applicable in X and favoring s, just in case there is no new inconsistency in the augmented case base $\Gamma \cup \{\langle X, r, s\rangle\}$ with respect to Γ.

And, given the generalized notion of a permitted rule, we can then say that the generalized reason model constrains the court to reach a decision on the basis of some applicable rule that is permitted according to the model.

To make Definition 2.7 less abstract, suppose that a court has to decide the situation $X_3 = \{f_1^\pi, f_2^\pi, f_1^\delta, f_2^\delta\}$ against the background of our earlier inconsistent case base $\Gamma_1 = \{c_1, c_2\}$. Is the court permitted to decide X_3 for the defendant on the basis of the rule $\{f_1^\delta\} \to \delta$? The answer is negative: Deciding X_3 in the suggested way would lead to the case $c_3 = \langle X_3, r_3, \delta\rangle$, where $r_3 = \{f_1^\delta\} \to \delta$, and to the priority $\{f_1^\pi, f_2^\pi\} <_{c_3} \{f_1^\delta\}$, which would conflict with the priority $\{f_1^\delta\} <_{c_1} \{f_1^\pi, f_2^\pi\}$ derived from the case c_1. Since the augmented case base $\Gamma_2 = \Gamma_1 \cup \{c_3\}$ contains c_1 and c_3, we would then have that $\{f_1^\delta\} <_{\Gamma_2} \{f_1^\pi, f_2^\pi\}$ and $\{f_1^\pi, f_2^\pi\} <_{\Gamma_2} \{f_1^\delta\}$. The case base Γ_2 would thus be inconsistent. Crucially, according to the generalized reason model notion of constraint, this would not be a problem if the inconsistency consisting of the reasons $\{f_1^\pi, f_2^\pi\}$ and $\{f_1^\delta\}$ were not new with respect to Γ_1—if, that is, we already had that $\{f_1^\delta\} <_{\Gamma_1} \{f_1^\pi, f_2^\pi\}$ and $\{f_1^\pi, f_2^\pi\} <_{\Gamma_1} \{f_1^\delta\}$. But this is not the case: Although the former priority holds because $\{f_1^\delta\} <_{c_1} \{f_1^\pi, f_2^\pi\}$ and c_1 belongs to Γ_1, the latter priority does not hold. In fact, among the cases belonging to Γ_1, only c_2 could support a priority of $\{f_1^\delta\}$ over $\{f_1^\pi, f_2^\pi\}$; yet, since $\{f_1^\pi, f_2^\pi\}$ does not hold in the fact situation X_2 decided in c_2, we do not have that $\{f_1^\pi, f_2^\pi\} <_{c_2} \{f_1^\delta\}$.

Does this mean that the court is not permitted to decide X_3 for δ at all? Well, No: For instance, by deciding X_3 for δ on the basis of the rule $\{f_1^\delta, f_2^\delta\} \to \delta$, the court would extend the case base Γ_1 with the case $c_4 = \langle X_4, r_4, \delta\rangle$, where $X_4 = X_3$ and $r_4 = \{f_1^\delta, f_2^\delta\} \to \delta$. This case, in turn, would lead to the priority ordering $<_{c_4}$ according to which every reason for the defendant that is at least as strong as $\{f_1^\delta, f_2^\delta\}$ has higher priority than each of the reasons $\{f_1^\pi\}$, $\{f_2^\pi\}$, and $\{f_1^\pi, f_2^\pi\}$ holding in X_4. Now, the only case in the augmented case base $\Gamma_3 = \Gamma_1 \cup \{c_4\}$ that could induce a priority ordering that conflicts with $<_{c_4}$ is the case c_1. Yet, the priority ordering $<_{c_1}$ does not conflict with $<_{c_4}$ because no reason that is at least as strong as $\{f_1^\delta, f_2^\delta\}$ holds in the fact situation X_1 decided in c_1, and so, if U is one such reason, we have that $\{f_1^\pi\} <_{c_4} U$, that $\{f_2^\pi\} <_{c_4} U$, and that $\{f_1^\pi, f_2^\pi\} <_{c_4} U$, but not that $U <_{c_1} \{f_1^\pi\}$, that $U <_{c_1} \{f_2^\pi\}$, or that $U <_{c_1} \{f_1^\pi, f_2^\pi\}$.[3] Since it would not introduce any new inconsistencies, deciding X_3 for δ on the basis of the rule r_4 is thus permitted.[4]

To conclude this section, it is immediate to verify that Definition 2.7 is indeed a generalization of the reason model notion of constraint:

Observation 2.8 *Let Γ be a consistent case base, X a new fact situation confronting the court, and r a rule applicable in X and favoring the side s. Then, there is no new inconsistency in the augmented case base $\Gamma \cup \{\langle X, r, s\rangle\}$ with respect to Γ if and only if the augmented case base $\Gamma \cup \{\langle X, r, s\rangle\}$ is consistent.*

Observation 2.8 tells us that, in the context of a consistent case base, a rule applicable to a new fact situation facing the court is permissible in the sense of the generalized reason model notion of constraint just in case it is permissible in the sense of the reason model notion of constraint.

[3] There is another reason why the priority $U <_{c_1} \{f_2^\pi\}$ does not hold, namely that the reason $\{f_2^\pi\}$ is not as strong as the premise of the rule of the case c_1, i.e., $\{f_1^\pi\}$.

[4] The reader can use arguments analogous to those presented in the last two paragraphs to verify that, against the background of Γ_1, the court is permitted to decide X_3 for π on the basis of the rule $\{f_1^\pi, f_2^\pi\} \to \pi$ but not on the basis of the rule $\{f_1^\pi\} \to \pi$.

3 Deontic logic

Observation 2.8 supports the idea that the notion of constraint set out in Definition 2.7 is a natural generalization of the reason model notion of constraint. But how exactly does the generalized notion work when the background case base is inconsistent? I start by exploring this question from a logical perspective: assuming the generalized notion, I aim, first, to define what it means to say that it follows from a possibly inconsistent case base that a decision for a side is obligatory or permitted and, second, to study the logic of constraint underlying the resulting notions of permission and obligation.

For the first task, I follow an idea from [13, Sect. 1.2.4]. Recall that Definition 2.7 characterizes the rules that, in the context of a certain case base, a court is permitted to use to justify its decisions in a particular fact situation. Given this notion, we can say that *it follows from a case base that it is permissible to decide a fact situation for a side* whenever there is a rule applicable to that fact situation that is permitted in the context of that case base and supports that side. We can also say that *it follows from a case base that it is obligatory to decide a fact situation for a side* whenever all rules applicable to that fact situation that are permitted in the context of that case base support that side. To state this formally, let $\Gamma \mathrel{|\!\sim} P_X(s)$ mean that it follows from Γ that deciding X for s is permitted according to the generalized reason model and $\Gamma \mathrel{|\!\sim} O_X(s)$ mean that it follows from Γ that deciding X for s is obligatory according to the generalized reason model—in the following, I will also use $\Gamma \not\mathrel{|\!\sim} P_X(s)$ to indicate that it does not follow from Γ that deciding X for s is permitted according to the generalized reason model, and similarly for $\Gamma \not\mathrel{|\!\sim} O_X(s)$. Then, $\Gamma \mathrel{|\!\sim} P_X(s)$ and $\Gamma \mathrel{|\!\sim} O_X(s)$ are defined as follows:

Definition 3.1 [Deontic operators] Let Γ be a case base and X a new fact situation confronting the court. Then, $\Gamma \mathrel{|\!\sim} P_X(s)$ holds if and only if, against the background of Γ, there is a rule applicable in X that is permitted by the generalized reason model and favors s, and $\Gamma \mathrel{|\!\sim} O_X(s)$ holds if and only if, against the background of Γ, every rule applicable in X that is permitted by the reason model favors s.

To illustrate, we have seen in the previous section that, against the background of the case base Γ_1, the rule $\{f_1^\delta, f_2^\delta\} \rightarrow \delta$ applicable in the fact situation X_3 and favoring the defendant is permitted by the generalized reason model. And, as mentioned in Footnote 4, the rule $\{f_1^\pi, f_2^\pi\} \rightarrow \pi$ applicable in X_3 and favoring the plaintiff is also permitted by the generalized reason model. So, both $\Gamma_1 \mathrel{|\!\sim} P_{X_3}(\delta)$ and $\Gamma_1 \mathrel{|\!\sim} P_{X_3}(\pi)$ hold, while neither $\Gamma_1 \mathrel{|\!\sim} O_{X_3}(\delta)$ nor $\Gamma_1 \mathrel{|\!\sim} O_{X_3}(\pi)$ hold.

Turning now to the logic of constraint underlying the introduced notions of permission and obligation, it immediately follows from our definitions that a court ought to decide a given fact situation for a side if and only if it is not permitted to decide that fact situation for the opposite side and, conversely, it follows from a case base that a court is permitted to decide a given fact situation for a side if and only if it does not follow from that case base that the court ought to decide that fact situation for the opposite side:

Observation 3.2 *Let Γ be a case base and X a new fact situation confronting the court. Then the following hold:*

1. $\Gamma \mid\!\sim O_X(s)$ *if and only if* $\Gamma \not\mid\!\sim P_X(\overline{s})$*;*
2. $\Gamma \mid\!\sim P_X(s)$ *if and only if* $\Gamma \not\mid\!\sim O_X(\overline{s})$*.*

Observation 3.2 tells us that the deontic operators introduced above are interdefinable in the usual way. A key question is whether, as it is often assumed in deontic logic, they also exclude the possibility of conflicting obligations: Can we exclude that, in the context of an inconsistent case base, a court is required to decide for the side s and also required to decide for the opposite side $\overline{s}$? The question is not trivial because Definition 3.1 mirrors the semantics of standard deontic logics and, in standard deontic logics, inconsistent normative information does give rise to contradictory requirements. [5] Now, in our case, the only situation in which both $\Gamma \mid\!\sim O_X(s)$ and $\Gamma \mid\!\sim O_X(\overline{s})$ would hold is when no rule applicable in X is permitted in the context of Γ. Fortunately, it turns out that this situation can be excluded: no matter whether Γ is inconsistent or which factors are present in X, there is a rule that the court is permitted to use to decide X. This result is important because it guarantees that the generalized reason model can sensibly guide a court's decision in every fact situation.

Observation 3.3 *Let Γ be a case base and X a new fact situation confronting the court. Then there exists some rule r applicable in X such that there is no new inconsistency in the augmented case base $\Gamma \cup \{\langle X, r, outcome(r)\rangle\}$ with respect to Γ.* [6]

An immediate consequence of Observation 3.3 is that, regardless of whether the background case base is inconsistent, the court will never be subject to contradictory requirements; in addition, in any situation, the court will be either required to decide for a side, or required to decide for the opposite side, or permitted to decide for either side:

Observation 3.4 *It is never the case that both $\Gamma \mid\!\sim O_X(s)$ and $\Gamma \mid\!\sim O_X(\overline{s})$ hold. In addition, it is always the case that exactly one of the following holds: either $\Gamma \mid\!\sim O_X(s)$, or $\Gamma \mid\!\sim O_X(\overline{s})$, or both $\Gamma \mid\!\sim P_X(s)$ and $\Gamma \mid\!\sim P_X(\overline{s})$.*

We can thus conclude that possibly inconsistent case bases support a natural, conflict-free deontic logic.

[5] Definition 3.1 mirrors the semantics of standard deontic logics in the sense that the deontic operators of permission and obligation are interpreted, respectively, as existential and universal quantifiers over a set of "permissible" or "ideal" entities, where the entities in question are rules in our case and possible worlds in the case of standard deontic logic. There is also a deeper connection between the two semantics, based on the possibility of constructing a Kripke model from case bases and fact situations. For reasons of space, I cannot present the details here.

[6] Observation 3.3 can be proved by adapting the proof of Observation 1 in [13, App. A.2]. The proof, which is not particularly complicated, is omitted for reasons of space.

4 Reformulating the generalized reason model

The previous conclusion—that, given the generalized reason model notion of constraint, possibly inconsistent case bases support a conflict-free deontic logic—is highly desirable on the one hand but still puzzling on the other: What is, exactly, the mechanism through which the generalized reason model extracts consistent requirements from an inconsistent case base? I will answer this question by proposing an illuminating reformulation of the generalized reason model notion of constraint. But, to do this, I first need to introduce a new distinction between cases supporting a potential decision and cases conflicting with it.

The basic idea behind the two new notions is simple and in line with a common understanding of legal arguments: Suppose that a court, facing a new fact situation X, wants to determine whether, against the background of a case base Γ, it is permissible to decide X for the side s on the basis of the rule r. The first thing that the court would do in this situation is to take the potential decision $c = \langle X, r, s \rangle$ and see if Γ contains precedent cases that support that decision or cases that conflict with that decision; depending on its finding, the court would then either retain or disregard the potential decision in question.

What does it mean, in the framework of the generalized reason model, that a precedent case $c_i = \langle X_i, r_i, s_i \rangle$ supports the potential decision $c = \langle X, r, s \rangle$? I will take this to mean two things: First, the precedent case c_i and the potential decision c have the same outcome—that is, $s_i = s$. And, second, the potential decision c is at least as strong for the winning side s as the precedent case c_i. But when is it the case that the potential decision c at least as strong for the winning side s as the precedent case c_i? Well, when it satisfies two conditions: First, the reason justifying the potential decision c is at least as strong as the reason justifying the precedent decision c_i—that is, $\mathit{premise}(r_i) \subseteq \mathit{premise}(r)$. Second, the strongest reason for $\overline{s}$ holding in the situation X_i decided in the precedent case c_i is at least strong as the strongest reason for the losing side $\overline{s}$ holding in the new situation X—that is, $X^{\overline{s}} \subseteq X_i^{\overline{s}}$. In other words, a precedent case decided for the side s supports a potential decision for s when the potential decision presents a justification for s that is at least as strong as that presented in the precedent case, in the context of a fact situation that includes reasons for $\overline{s}$ that are weaker than those included in the precedent fact situation. [7] Formally, the set $\mathit{supporting}_\Gamma(c)$ of cases from Γ that support the potential

[7] The notion of a supported case proposed here is a generalization of the notion of an *a fortiori* case discussed in [2] and formalized in [11]. This notion is defined in the context of a model of precedential constraint, called the result model, that can be obtained from the reason model by adding the requirement that the only reasons a court can use to justify its decision regarding a fact situation X are either X^s or $X^{\overline{s}}$. In this framework, a potential decision $c = \langle X, X^s \to s, s \rangle$ is said to be *a fortiori* given a precedent case $c_i = \langle X_i, r_i, s_i \rangle$ such that $s_i = s$ just in case $X_i^s \subseteq X^s$ and $X^{\overline{s}} \subseteq X_i^{\overline{s}}$. According to Prakken and Ratsma [18], the precedents that make a potential decision *a fortiori* are the best precedents to cite in support of that decision. The notion of support defined here can thus also be viewed as a generalization of Prakken's and Ratsma's notion of best precedent to cite. For the reader familiar with the literature in AI and Law, it might be worth noticing that the latter notion does not coincide with the familiar notion of best precedent to cite form HYPO [3].

decision c is thus defined as follows:

Definition 4.1 [Supporting cases.] Let Γ be a case base and $c = \langle X, r, s\rangle$ a new potential decision. Then,

$$\begin{aligned} \mathit{supporting}_\Gamma(c) = \{c_i = \langle X_i, r_i, s_i\rangle \in \Gamma : {} & (1)\ s_i = s, \text{ and} \\ & (2)\ \mathit{premise}(r_i) \subseteq \mathit{premise}(r), \text{ and} \\ & (3)\ X^{\overline{s}} \subseteq X_i^{\overline{s}}\} \end{aligned}$$

To illustrate the definition, let us go back to the earlier inconsistent case base $\Gamma_1 = \{c_1, c_2\}$ and to the fact situation $X_3 = \{f_1^\pi, f_2^\pi, f_1^\delta, f_2^\delta\}$. We can easily verify that the precedent case c_2 does not support the potential decision $c_3 = \langle X_3, r_3, \delta\rangle$, where $r_3 = \{f_1^\delta\} \to \delta$. In fact, even if, in accordance with condition 2, the precedent c_2 and the potential decision c_3 justify a decision for the defendant on the basis of the same reason, the strongest reason for the plaintiff that holds in X_3 (i.e., $\{f_1^\pi, f_2^\pi\}$) is stronger than the strongest reason for the plaintiff that holds in $\mathit{facts}(c_2)$ (i.e., $\{f_1^\pi\}$), against condition 3. Since c_2 is the only precedent in Γ_2 that was decided for the defendant, this means that $\mathit{supporting}_{\Gamma_1}(c_3) = \varnothing$. Let me also note (for later) that a similar argument shows that, where $c_4 = \langle X_4, r_4, \delta\rangle$ with $X_4 = X_3$ and $r_4 = \{f_1^\delta, f_2^\delta\} \to \delta$, we also have that $\mathit{supporting}_{\Gamma_2}(c_4) = \varnothing$.

Before moving on to the notion of a conflicting case, it is crucial to observe that, when a potential decision is supported by a precedent case, the information about the priority of reasons derived from that decision can already be derived from the supporting precedent. Another way of thinking of a supporting case is then as a precedent case that supports the priority of reasons derived from the supported potential decision:

Observation 4.2 *Let Γ be a case base, $c_i = \langle X_i, r_i, s\rangle$ a case in Γ, and $c = \langle X, r, s\rangle$ a new potential decision. Then c_i belongs to $\mathit{supporting}_\Gamma(c)$ if and only if, for every pair of opposing reasons U and V, if $U <_c V$, then $U <_{c_i} V$.*[8]

It follows immediately from Observation 4.2 that, when we augment a case base Γ with a decision c that is supported by a case from Γ, the priority ordering $<_{\Gamma \cup \{c\}}$ induced by the augmented case base $\Gamma \cup \{c\}$ is just the same as the priority ordering $<_\Gamma$ induced by the initial case base Γ—that is, the potential decision c does not tell us anything new about the priority of reasons.

Now, when a court evaluates a potential decision, it considers not only whether some precedent cases support that decision but also whether some precedent cases conflict with it. In the framework of the generalized reason model, it is natural to say that a precedent $c_j = \langle X_j, r_j, s_j\rangle$ conflicts with a potential decision $c = \langle X, r, s\rangle$ when three conditions obtain: First, the precedent c_j and the potential decision c have different outcomes—that is, $s_j = \overline{s}$. Second, the reason justifying the precedent decision for $\overline{s}$ holds in the fact sit-

[8] Both directions of Observation 4.2 are an immediate consequence of Definition 2.1, Definition 4.1, and, for the right-to-left direction, the fact that $X^{\overline{s}} <_c \mathit{premise}(r)$. The details of the proof are left to the reader.

uation the potential decision is about—that is, $\mathit{premise}(r_j) \subseteq X$. And, finally, according to the priority ordering derived from the precedent case, the reason justifying the potential decision for s has lower priority than the reason justifying the precedent decision for $\overline{s}$—that is, $\mathit{premise}(r) <_{c_j} \mathit{premise}(r_j)$. Intuitively, these three conditions can be taken to say that, according to the precedent c_j, the new fact situation X should be decided for $\overline{s}$ on the basis of $\mathit{premise}(r_j)$ rather than for s on the basis of $\mathit{premise}(r)$.[9] Since, by Definition 2.1, the third condition is equivalent to the requirement that the reason $\mathit{premise}(r)$ holds in X_j, we can define the set $\mathit{conflicting}_\Gamma(c)$ of cases from Γ that conflict with the potential decision c as follows:

Definition 4.3 [Conflicting cases.] Let Γ be a case base and $c = \langle X, r, s\rangle$ a new potential decision. Then,

$$\begin{aligned} \mathit{conflicting}_\Gamma(c) = \{c_j = \langle X_j, r_j, s_j\rangle \in \Gamma : \; & (1)\ s_j = \overline{s},\ \text{and} \\ & (2)\ \mathit{premise}(r_j) \subseteq X,\ \text{and} \\ & (3)\ \mathit{premise}(r) \subseteq X_j\} \end{aligned}$$

Going back to our previous example, it is not difficult to see that the precedent case c_1 conflicts with the potential decision c_3 because the reason justifying a decision for the plaintiff in c_1, i.e., the reason $\{f_1^\pi\}$, holds in $X_3 = \{f_1^\pi, f_2^\pi, f_1^\delta, f_2^\delta\}$ and, in turn, the reason justifying a decision for the defendant in c_3, i.e., $\{f_1^\delta\}$, holds in $X_1 = \{f_1^\pi, f_2^\pi, f_1^\delta\}$. Hence, $c_1 \in \mathit{conflicting}_{\Gamma_1}(c_3)$. On the other hand, the case c_1 does not conflict with the potential decision $c_4 = \langle X_4, r_4, \delta\rangle$, because the reason justifying a decision for the defendant in c_4, i.e., the reason $\{f_1^\delta, f_2^\delta\}$, does not hold in X_1. Since c_1 is the only case in Γ_1 that could conflict with c_4, we then have that $\mathit{conflicting}_{\Gamma_1}(c_4) = \varnothing$.

At this point, a natural question arises: In discussing Observation 4.2, I suggested that a supporting case can be thought of as a case supporting the priority of reasons derived from the supported potential decision. Can we think of a conflicting case as a case that conflicts with the priority of reasons derived from the conflicted potential decision? The answer is Yes: when a precedent case conflicts with a new potential decision, there is a priority relation between two opposing reasons that the precedent case and the potential decision disagree about.

[9] In the AI and Law literature, legal arguments are typically analysed as having a “three-ply” structure, where, first, the proponent presents a precedent case supporting their claim, then the opponent responds by either distinguishing the precedent case from the focus case advanced by the proponent or by presenting another precedent case as a counterexample, and finally the proponent attempts to rebut the opponent’s response (for a recent overview of models of legal arguments in AI and Law see [4]). In this context, it is natural to think of conflicting cases as counterexamples. I did not choose this terminology, first, to avoid confusion with other notions of counterexample present in the literature and, second, to avoid suggesting that I am trying to model three-ply arguments—even if supporting and conflicting cases may be used in a legal argument, there is no explicit representation of legal arguments in the (generalized) reason model.

Observation 4.4 *Let Γ be a case base, $c_j = \langle X_j, r_j, \overline{s}\rangle$ a case in Γ, and $c = \langle X, r, s\rangle$ a new potential decision. Then c_j belongs to conflicting$_\Gamma(c)$ if and only if there is a pair of opposing reasons U and V such that $U <_c V$ and $V <_{c_j} U$.*

In fact, the reader can easily verify that Observation 4.4 holds, for instance, when $U = \mathit{premise}(r_j)$ and $V = \mathit{premise}(r)$ or when $U = X^{\overline{s}}$ and $V = \mathit{premise}(r)$. So, the precedent c_j conflicts with the potential decision c whenever c_j and c disagree about the relative importance of their respective justifications or, equivalently, whenever they disagree about whether the reason *premise*(r) is in fact important enough to justify a decision for s in spite of the presence of the reason $X^{\overline{s}}$.

We are now ready to reformulate the generalized reason model notion of constraint. Recall that, according to this notion, the court is permitted, against the background of a case base Γ, to decide a fact situation X for the side s on the basis of the rule r just in case there is no new inconsistency in the augmented case base $\Gamma \cup \{\langle X, r, s\rangle\}$ with respect to Γ, where a new inconsistency is an inconsistency in the augmented case base that was not an inconsistency in the initial case base. Let us now ask: When is it the case that extending Γ with a potential decision of the form $\langle X, r, s\rangle$ leads to an inconsistency? Well, given Observation 4.4, the answer is simply that this happens when there is a case in Γ that conflicts with $\langle X, r, s\rangle$. But when is it the case that, in addition, the inconsistency in question is new? Well, given our Observation 4.2, the inconsistency is new when there is no precedent case in Γ that supports the potential decision $\langle X, r, s\rangle$. More precisely:

Observation 4.5 *Let Γ be a case base and $c = \langle X, r, s\rangle$ a new potential decision considered by the court. Then, there is a pair of opposing reasons U and V such that it is the case that $U <_{\Gamma\cup\{c\}} V$ and $V <_{\Gamma\cup\{c\}} U$ but not the case that $U <_\Gamma V$ and $V <_\Gamma U$ if and only if conflicting$_\Gamma(c) \neq \varnothing$ and supporting$_\Gamma(c) = \varnothing$.*[10]

So, according to Observation 4.5, the court is *not* permitted to decide a fact situation X for the side s on the basis of the rule r against the background of Γ whenever there is a precedent case in Γ that conflicts with this potential decision but no precedent case in Γ that supports it. If we go back one more time to the example of the case base Γ_1 and the new fact situation X_3, we can now see that the potential decision c_3 is not permitted because, as we have seen above, the precedent case c_1 conflicts with it while no precedent case supports it. On the other hand, the potential decision c_4 is permitted because, even if no precedent case supports it, no precedent case conflicts with it either.

[10] The left-to-right direction of Observation 4.5 follows immediately from Observation 4.2 and Observation 4.4. For the left-to-right direction, the reader can verify that, if *conflicting*$_\Gamma(c) \neq \varnothing$ and *supporting*$_\Gamma(c) = \varnothing$, then the reasons *premise*(r) and $X^{\overline{s}}$ form a new inconsistency in $\Gamma \cup \{c\}$ with respect to Γ.

5 Back to deontic logic

Let us now go back to the notion of obligation set out in Definition 3.1. Observation 3.2 told us that, according to this notion, a court ought to decide a new fact situation X for the side s against the background of a case base Γ just in case it is *not* permitted to decide X for $\overline{s}$ against the background of Γ:

$$\Gamma \mid\!\sim O_X(s) \text{ iff } \Gamma \mid\!\not\sim P_X(\overline{s}) \tag{1}$$

In turn, by Definition 3.1, the right-hand side of (1) holds just in case, against the background of Γ, there is no rule applicable in X and favoring $\overline{s}$ that is permitted by the generalized reason model. Given our reformulation of the generalized reason model in Observation 4.5, this means that, for every potential decision $c = \langle X, r, \overline{s}\rangle$, there is a case in Γ that conflicts with c but no case in Γ that supports c. By distributing the universal quantifier we then get:

$$\begin{aligned}\Gamma \mid\!\sim O_X(s) \text{ iff, for all } c = \langle X, r, \overline{s}\rangle &: \textit{conflicting}_\Gamma(c) \neq \varnothing \text{ and}\\ \text{for all } c = \langle X, r, \overline{s}\rangle &: \textit{supporting}_\Gamma(c) = \varnothing\end{aligned} \tag{2}$$

Conveniently, it turns out that, for every potential decision of the form $c = \langle X, r, \overline{s}\rangle$, there is a case in Γ that conflicts with c just in case there is a case in Γ that conflicts with the potential decision

$$c_{X^{\overline{s}}} = \left\langle X, X^{\overline{s}} \to \overline{s}, \overline{s}\right\rangle.\ ^{11}$$

We can then replace the biconditional (2) with the following biconditional:

$$\begin{aligned}\Gamma \mid\!\sim O_X(s) \text{ iff } &\textit{conflicting}_\Gamma(c_{X^{\overline{s}}}) \neq \varnothing \text{ and}\\ &\text{for all } c = \langle X, r, \overline{s}\rangle : \textit{supporting}_\Gamma(c) = \varnothing\end{aligned} \tag{3}$$

It also turns out that, for every potential decision of the form $c = \langle X, r, \overline{s}\rangle$, there is no case in Γ that supports c just in case there is no case in Γ that supports the potential decision $c_{X^{\overline{s}}}$.[12] We can then replace the biconditional (3) with:

$$\begin{aligned}\Gamma \mid\!\sim O_X(s) \text{ iff } &\textit{conflicting}_\Gamma(c_{X^{\overline{s}}}) \neq \varnothing \text{ and}\\ &\text{iff } \textit{supporting}_\Gamma(c_{X^{\overline{s}}}) = \varnothing\end{aligned} \tag{4}$$

One final transformation and we are done: it follows immediately from our definitions of a conflicting and of a supporting case that a precedent case conflicts with the potential decision $c_{X^{\overline{s}}}$ if and only if it supports the potential decision

$$c_{X^s} = \langle X, X^s \to s, s\rangle .$$

[11] This fact is an immediate consequence of Definition 2.1 and Definition 4.3.

[12] This fact is an immediate consequence of Definition 2.1 and Definition 4.1.

That is: $\mathit{conflicting}_\Gamma(c_{X^{\overline{s}}}) = \mathit{supporting}_\Gamma(c_{X^s})$. We can thus replace the biconditional (4) with:

$$\begin{aligned}\Gamma \mathrel{|\!\sim} O_X(s) \text{ iff } & \mathit{supporting}_\Gamma(c_{X^s}) \neq \varnothing \text{ and} \\ \text{iff } & \mathit{supporting}_\Gamma(c_{X^{\overline{s}}}) = \varnothing\end{aligned} \tag{5}$$

And, of course, form the previous biconditional and Observation 3.2, we also get:

$$\begin{aligned}\Gamma \mathrel{|\!\sim} P_X(s) \text{ iff } & \Gamma \not\mathrel{|\!\sim} O_X(\overline{s}) \\ \text{iff either } & \mathit{supporting}_\Gamma(c_{X^{\overline{s}}}) = \varnothing \\ \text{or } & \mathit{supporting}_\Gamma(c_{X^s}) \neq \varnothing\end{aligned} \tag{6}$$

At this point, we are in a position to understand how it is, exactly, that the generalized reason model extracts consistent obligations from inconsistent case bases. In order to determine whether deciding X for s is obligatory, the model does the following: First, it restricts the set of potential decisions concerning X to the set $\{c_{X^s}, c_{X^{\overline{s}}}\}$. Then, it inspects the background case base Γ and determines, for each of the two potential decisions, whether there are cases in Γ that support it. Depending on the result of the analysis, the model reaches a conclusion as shown in the following table:

	$\mathit{supporting}_\Gamma(c_{X^s}) = \varnothing$	$\mathit{supporting}_\Gamma(c_{X^s}) \neq \varnothing$
$\mathit{supporting}_\Gamma(c_{X^{\overline{s}}}) = \varnothing$	$^{(A)}$ $\Gamma \mathrel{|\!\sim} P_X(\overline{s})$ and $\Gamma \mathrel{|\!\sim} P_X(s)$	$^{(B)}$ $\Gamma \not\mathrel{|\!\sim} P_X(\overline{s})$ and $\Gamma \mathrel{|\!\sim} O_X(s)$
$\mathit{supporting}_\Gamma(c_{X^{\overline{s}}}) \neq \varnothing$	$^{(C)}$ $\Gamma \mathrel{|\!\sim} O_X(\overline{s})$ and $\Gamma \not\mathrel{|\!\sim} P_X(s)$	$^{(D)}$ $\Gamma \mathrel{|\!\sim} P_X(\overline{s})$ and $\Gamma \mathrel{|\!\sim} P_X(s)$

Now, recall that a supporting case can be viewed as a case that supports the priority of reasons derived from the supported case. This notion of support has, in fact, the strength of an entailment relation: every priority that can be derived from the supported case can already be derived from the supporting case. This might lead us to think that the presence of a supporting relation generates a sort of requirement—the supporting case requires, in a sense, that the court adopt the priority ordering derived from the supported case. [13] If we assume this interpretation, then it makes perfect sense that, in the cases (A), (B), and (C) from the table above, the model draws the conclusion that it does: In the case (A), there is no precedent case requiring either a decision for s or a decision for $\overline{s}$, so it is permissible to decide for either side. On the other hand, in the cases (B) and (C), there is a precedent requiring a decision for one of the two sides but no precedent requiring a decision for the other, so it is obligatory to satisfy the only existing requirement. The problem, of course, arises in the case (D), where there is both a precedent requiring a decision for s and a precedent requiring a decision for $\overline{s}$. Here is where the biconditional (5) reveals that the generalized reason model simply builds consistency into the

[13] In fact, in the context of the result model (see Footnote 7), a supporting case is sometimes said to "control" [10] or "force" [18] the supported potential decision; in addition, the presence of a supporting case in the background case base is what grounds a requirement to decide for the side that won the supporting case (see [13], Definition 21).

notion of obligation: when two precedents support conflicting decisions, the model basically ignores the conflicting requirements and concludes that either decision is permissible.

6 Conclusion

I investigated a generalization of the reason model notion of precedential constraint that can be used to address the question *How does an inconsistent set of precedent cases constrain future decisions?* I started to explore this question in Section 3 from a logical perspective. I have shown, first, that the model supports an interesting deontic logic, where consistent obligations can be derived from inconsistent case bases. I then provided an explanation of this surprising result in two steps: In Section 4, I proposed a reformulation of the generalized reason model notion of constraint in terms of cases that support a new potential decision and cases that conflict with it. In Section 5, I then observed that it follows from the proposed reformulation that a court is obliged to decide a fact situation for a side just in case the background case base contains a case that supports the strongest potential decision for that side but no case that supports the strongest potential decision for the opposite side. I argued that this shows that the generalized reason model builds consistency into the notion of obligation. I close by simply mentioning two open issues:

First, the generalized reason model starts with a factor-based representation of legal cases, where each factor is binary in the sense that it fully favors either the plaintiff or the defendant. But, in real cases, it is not unusual that some facts favor either one of the two sides, not fully, but only to a certain extent. For instance, going back to the example at the beginning of the paper, a relevant consideration to decide whether a student should be granted an extension on an assignment might be whether the student asked for extensions before. This consideration is best represented not as a binary, all-or-nothing factor, but as a multi-valued factor, where the different values are the numbers of times the student asked for an extension and greater values progressively weakens the student's case. As mentioned in the introduction, several scholars have proposed ways to refine the reason model to account for multi-valued factors of this sort [12,17,21]. A natural question is then whether these refinements of the reason model can be generalized to apply to inconsistent case bases along the lines proposed in this paper and, if so, whether the results obtained in Sections 3 to 5 still hold.

A different issue concerns the possibility of extracting, from an inconsistent case base, not only obligations and permissions, but also explanations for particular decisions. Recent work has shown how a consistent case base can be mapped into default logic [13], logic programming [8], or an abstract argumentation framework [9,18]. The latter formalisms can then be used to generate so-called argumentative explanations of decisions made against the background of the case base under consideration. The question here would be whether the generalized reason model could help us generalize these mappings to inconsistent case bases as well and, if so, whether the generated explanations would still be sensible.

Acknowledgement

This research is partly supported by the project "Responsible Artificial Agency: A Logical Perspective," funded by the Research Priority Area Human(e) AI at the University of Amsterdam.

References

[1] Aleven, V., "Teaching Case-Based Argumentation Through a Model and Examples," Ph.D. thesis, PhD Thesis, Intelligent Systems Program, University of Pittsburgh (1997).

[2] Alexander, L., *Constrained by precedent*, Southern California Law Review **63** (1989), pp. 1–64.

[3] Ashley, K., "Modeling Legal Argument: Reasoning with Cases and Hypotheticals," The MIT Press, 1990.

[4] Atkinson, K. and T. Bench-Capon, *Argumentation schemes in AI and Law*, Argument & Computation **12** (2021).

[5] Bench-Capon, T., *Some observations on modelling case based reasoning with formal argument models*, in: *Proceedings of the Sixth International Conference on Artificial Intelligence and Law (ICAIL-99)* (1999), pp. 36–42.

[6] Bench-Capon, T. and K. Atkinson, *Precedential constraint: The role of issues*, in: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (ICAIL 2021)* (2021), pp. 12–21.

[7] Canavotto, I. and J. Horty, *Piecemeal knowledge acquisition for computational normative reasoning*, in: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'2022)* (2022), pp. 171–180.

[8] Cocarascu, O., K. Čyras and F. Toni, *Explanatory predictions with artificial neural networks and argumentation*, in: *Proceedings of the IJCAI/ECAI-2018 Workshop on Explainable Artificial Intelligence*, 2018, pp. 26–32.

[9] Čyras, K., K. Satoh and F. Toni, *Explanation for case-based reasoning via abstract argumentation*, in: *Computational Models of Argument. Proceedings of COMMA 2016* (2016), pp. 26–32.

[10] Horty, J., *The result model of precedent*, Legal Theory **10** (2004), pp. 19–31.

[11] Horty, J., *Rules and reasons in the theory of precedent*, Legal Theory **17** (2011), pp. 1–33.

[12] Horty, J., *Reasoning with dimensions and magnitudes*, Artificial Intelligence and Law **27** (2019), pp. 307–345.

[13] Horty, J., "*The Logic of Precedent: Constraint and Freedom in Common Law Reasoning*," 20xx, forthcoming with Cambridge University Press. Available at `http://www.horty.umiacs.io/articles/2022-7-15-logic-precedent.pdf`.

[14] Horty, J. and T. Bench-Capon, *A factor-based definition of precedential constraint*, Artificial Intelligence and Law **20** (2012), pp. 181–214.

[15] Lamond, G., *Do precedents create rules?*, Legal Theory **11** (2005), pp. 1–26.

[16] Peters, J., H. Prakken and F. Bex, *Justification derived from inconsistent case bases using authoritativeness*, in: *Proceedings of the First International Workshop on Argumentation for eXplainable AI (ArgXAI)* (2022).

[17] Prakken, H., *A formal analysis of some factor- and precedent-based accounts of precedential constraint*, Artificial Intelligence and Law **29** (2021), pp. 559–585.

[18] Prakken, H. and R. Ratsma, *A top-level model of case-based argumentation for explanation: Formalisation and experiments*, Argument & Computation **13** (2022), pp. 159–194.

[19] Prakken, H. and G. Sartor, *Modelling reasoning with precedents in a formal dialogue game*, Artificial Intelligence and Law **6** (1998), pp. 231–287.

[20] Rigoni, A., *An improved factor based approach to precedential constraint*, Artificial Intelligence and Law **23** (2015), pp. 133–160.

[21] Rigoni, A., *Representing dimensions within the reason model of precedent*, Artificial Intelligence and Law **26** (2018), pp. 1–22.

[22] van Woerkom, W., D. Grossi, H. Prakken and B. Verheij, *Landmarks in case-based reasoning: From theory to data*, in: *Proceedings of the First International Conference on Hybrid Human-Machine Intelligence* (2022).

What should I do and why?

Joris Hulstijn and Leon van der Torre

University of Luxembourg,
joris.hulstijn@uni.lu, leon.vandertorre@uni.lu

Keywords: explanation, deontic logic, questions, actions

There is a lot of interest in explainable AI [2,11].When a system takes decisions that affect people, they can demand an explanation of how the decision was derived, or a justification of why the decision is justified. Note that explanation and justification are related, but not the same [1]. The need for explanation or justification is more pressing, when the system makes legal decisions [3], or when the decision is based on social or ethical norms [5]. Requests for explanation and justification are typically made in a dialogue setting [5].

(1) A. You must do the dishes!
B. Why?
A. Because someone must do the dishes! And I don't have time.

What is the meaning of a why-question in a deontic context? What qualifies as an answer? We aim to provide a semantics of deontic explanations as answers to why-questions. Our method is to combine three types of semantics, based on a partition or equivalence relation over possible worlds: a logic of questions and answers by Groenendijk and Stokhof [8], a logic of obligations or norms, by Kratzer [6,10], and a logic of choice and action, by Horty [9]. A question indicates various possible answers, which correspond to a choice of actions, the outcomes of which are ordered by a normative preference order.

A's imperative in (1) presupposes that *B* has a choice to obey or not. In STIT-logic that choice structure can be modelled by two equivalence classes of histories: those of doing the dishes and those of not doing the dishes. All other differences between possible worlds are abstracted over. This choice structure turns out to correspond to the semantics of a question from *B*'s perspective: shall I do the dishes or not? The obligation means that there is a preference: worlds in which *B* does the dishes are strictly preferred over similar worlds, in which *B* does not do the dishes, at least according to *A* (Figure 1 left).

Now consider *B*'s why-question. Just like a who-question asks for persons, a why-question asks for reasons. What reasons qualify as an explanation in this context? Here, a reason is a proposition, that will help to reduce the context set, so the preferred alternative remains. But the why-question is not exactly

Fig. 1. Analogy between questions, obligations and STIT, contrasting alternatives

the choice! For A to order B to do the dishes, some felicity conditions must hold [4]. A dialogue context is generally underspecified. That triggers questions under discussion [7], whose answers are relevant. Are there rules about doing the dishes? Does A have authority over B? Who else is present?

Day to day explanations are supposed to be *contrastive*: they show why the outcome is to be preferred over some counterfactual alternatives [11]. A reason should resolve enough of the underspecification in the context, to persuade the other to select the preferred alternative.

(2) A. You must do the dishes!
B. Why $[\text{must}]_F$ I do the dishes?
B. Why must $[\text{I}]_F$ do the dishes?
B. Why must I do $[\text{the dishes}]_F$?

What are the possible alternatives? In example (2) alternatives are marked by focus, see [12]. The original why-question is ambiguous. Dialogue participants have the freedom to 'take-up' a why-question in different ways and negotiate sensitive issues, such as authority [4]. Here B chooses to answer by referring to a rule: someone must do the dishes. The underspecified nature of 'someone' triggers a question-under-discussion 'who' (Figure 1 middle). The rest of the answer rules out A. Only one alternative remains: B must do the dishes.

So, by combining linguistics and deontic logic, we develop the idea for a semantics of deontic explanations in a dialogue context. A deontic explanation is an answer to a question-under-discussion triggered by an underspecified part of the dialogue context. That answer reduces the context set to one of the preferred alternatives in the partition that is induced by the choice structure.

References

[1] Alvarez, M.: Reasons for Action: Justification, Motivation, Explanation (2017)

[2] Anjomshoae, S., Calvaresi, D., Najjar, A., Främling, K.: Explainable agents and robots: Results from a systematic literature review. In: (AAMAS 2019), pp. 1078–1088 (2019)

[3] Atkinson, K., Bench-Capon, T., Bollegala, D.: Explanation in AI and Law: Past, present and future. Artificial Intelligence **289**, 103387 (2020)

[4] Austin, J.L.: How to do things with words. Harvard University Press (1962)

[5] van Berkel, K., Straßer, C.: Reasoning with and about norms in logical argumentation. In: (COMMA 2022). vol. 353, pp. 332–343. IOS Press (2022)

[6] Gabbay, D., Horty, J., Parent, X., van der Meyden, R., van der Torre, L. (eds.): Handbook of Deontic Logic and Normative Systems. College Publications (2013)

[7] Ginzburg, J.: Resolving questions, i. Linguistics and Philosophy **18**, 459–527 (1995)

[8] Groenendijk, J., Stokhof, M.: Questions. In: Van Benthem, J., Ter Meulen, A. (eds.) Handbook of Logic and Language, pp. 1055–1124. North-Holland, Elsevier (1996)

[9] Horty, J.F.: Agency and Deontic Logic. Oxford University Press, New York (2001)

[10] Kratzer, A.: Modals and Conditionals. Oxford University Press (2012)

[11] Miller, T.: Explanation in Artificial Intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)

[12] Rooth, M.: A theory of focus interpretation. Natural Language Semantics **1**(1), 75–116 (1992)

Conceptual and Logical Analysis of the Right to Know

Réka Markovich [1]

University of Luxembourg

Olivier Roy

University of Bayreuth

Keywords: normative positions, rights, multi-modal logic, static and dynamic logics

We study the right to know and its logical characterization in the theory of normative positions. The theory of normative positions [3] stems from the work of the American legal theorist Hohfeld who, finding that the word 'right' was overused and referred to various concepts causing a terminological and conceptual confusion, differentiated between four atomic types of rights (claim-right, privilege, power, immunity) and their correlative duties (duty, no-claim, liability, disability)[1]. Despite Hohfeld's pursuit, the legal terminology has not changed much ever since, thus identifying the consequences of a right today still requires investigating which atomic normative position or their molecular combinations it refers to and analyzing it with an adequate formalism.

We investigate and formalize the possible meaning of the right to know using different (monadic and dyadic) deontic logics extending them with epistemic and (legal)-alethic modalities. We propose several plausible but non-equivalent formalizations of the 'right to know whether'. Our first set of formalizations capture the right to know as a conditional claim-right, i.e. that one has the right to know ϕ, given that ϕ is the case. We compare these formalizations in terms of how they fare with respect to the so-called Aqvist paradox [2]. We furthermore investigate the logical behavior of these claim rights with respect to detachment and detachment principles, which are central to both legal theory and deontic logic. We finally study the logical relationship between the different formalizations.

Our second set of formalizations captures the right to know as a legal power. In the theory of normative positions, the claim-right-duty and privilege-no-

[1] This work was supported by the Fonds National de la Recherche Luxembourg through the project Deontic Logic for Epistemic Rights (OPEN O20/14776480) and the project INTEGRAUTO (INTER/AUDACE/21/16695098).

claim pairs refer to the static, deontic aspects: duty is a directed obligation, a privilege is a relational variant of weak permission in standard deontic logic. The power-liability and the immunity-disability pairs refer, in contrast, to an agent's capacity to change the counterparty's normative positions, thus to dynamic normative positions. After proposing a concrete formalization of the right to know as a power, or more precisely the power to create a claim right, we comment on the potential of this formalization to capture claim rights that are conditionals of certain epistemic conditions.

All in all, the paper makes a conceptual rather than technical contribution: it maps the possibilities of understanding the right to know as a claim-right and as a power, and shows more generally the fruitfulness of the theory of normative positions to the understanding of epistemic rights The meta-logical and computational properties of the underlying logic are left for future work.

References

[1] Hohfeld, W. N., *Fundamental legal conceptions applied in judicial reasoning*, in: W. W. Cook, editor, *Fundamental Legal Conceptions Applied in Judicial Reasoning and Other Legal Essays*, New Haven: Yale University Press, 1923 pp. 23–64.

[2] Åqvist, L., *Good samaritans, contrary-to-duty imperatives, and epistemic obligations*, Nous **1** (1967), pp. 361–379.
URL http://www.jstor.org/stable/2214624

[3] Sergot, M., *Normative Positions*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, College Publications, 2013 pp. 353–406.

Reason-based detachment

Aleks Knoks[1]

University of Luxembourg
2, avenue de l'Université
L-4365 Esch-sur-Alzette, Luxembourg

Leendert van der Torre[1]

University of Luxembourg
2, avenue de l'Université
L-4365 Esch-sur-Alzette, Luxembourg

Keywords: detachment, principles, reasons, weighing.

When philosophers talk about normative matters—about what is right, obligatory, permitted, and so on—they tend to rely on the notion of *normative reasons*, understanding them as considerations that count in favor of or against actions (or attitudes). The notion has become a mainstay of practical philosophy, where it is routinely made use of in answering various normative and metanormative questions. This is taken to the extreme in the *reasons-first program* which holds, roughly, that the notion of a reason is basic, and that all other normative notions are to be analyzed in terms of it.[2] When discussing the interaction between reasons, philosophers often use such phrases as "the action supported on the balance of reasons" and "reasons for outweigh reasons against", inviting an image of *weight scales.* The simplest version of these normative scales is meant to work roughly as follows. The reasons speaking in favor of φ-ing go in one pan of the scales, the reasons against φ-ing go in the other. If the weight of the reasons in the first pan is greater than that of the reasons in the second, φ ought to be carried out. If the weight of the reasons in the second pan is greater, φ ought not to be carried out.[3]

Philosophers have explored various ideas about the exact workings of the normative weight scales, as well as some alternatives to them, but, with few

[1] This research has been supported by the Fonds National de la Recherche Luxembourg (FNR) through the project Deontic Logic for Epistemic Rights (OPEN O20/14776480). The authors also acknowledge the travel support from the same funding institution through the project INTEGRAUTO.

[2] The locus classicus here is Scanlon [11]. But see also, e.g., Parfit [9], Raz [10], and Schroeder [12].

[3] Cf., e.g., Dancy [1], Lord and Maguire [5], and Tucker [13].

exceptions, these investigations have been carried out informally, while the more formal investigations have focused on exploring particular models. [4] Our alternative proposal is to think of the weight scales as a kind of inference pattern: the titular *reason-based detachment.* We set up a general formal framework built around it and report on the (first) results of exploring it. [5] The underlying idea is to start with the general notion of *detachment systems*—which can be thought of as structures in which reason-based detachment is guaranteed to be valid—formulate various principles or properties that a given detachment system can satisfy, and explore different classes of such systems. For example, we define the principle called *Neutrality* which requires, roughly, that reasons of opposing polarity—reasons for and against—are treated equally, and the principle called *Fixed Value* which requires that a reason's polarity is the same in every context. Of particular interest is the class that we call *balancing operations*, since the detachment systems in it reflect some of the core features of the informal idea of weighing scales. We also define several concrete balancing operations and present a principle-based analysis of reason-based detachment.

References

[1] Dancy, J., "Ethics without Principles," Oxford University Press, 2004.

[2] Dietrich, F. and C. List, *A reason-based theory of rational choice*, Noûs **47(1)** (2013), pp. 104–34.

[3] Faroldi, F. and T. Protopopescu, *A hyperintensional logical framework for deontic reasons*, Logic Journal o the IGPL (Forthcoming).

[4] Horty, J., "Reasons as Defaults," Oxford University Press, 2012.

[5] Lord, E. and B. Maguire, *An opinionated guide to the weight of reasons*, in: E. Lord and B. Maguire, editors, *Weighing Reasons*, Oxford University Press, 2016 pp. 3–24.

[6] Makinson, D. and L. van der Torre, *Input/output logics*, Journal of Philosophical Logic **29** (2000), pp. 383–408.

[7] Makinson, D. and L. van der Torre, *Constraints for input/output logics*, Journal of Philosophical Logic **30** (2001), pp. 155–85.

[8] Parent, X. and L. van der Torre, *Input/output logic*, in: L. van der Torre, D. Gabbay, J. Horty and R. van der Meyden, editors, *Handbook of Deontic Logic*, College Publications, 2013 .

[9] Parfit, D., "On What Matters," Oxford University Press, 2011.

[10] Raz, J., "Practical reason and norms," Oxford University Press, 1990.

[11] Scanlon, T. M., "What We Owe to Each Other," Cambridge, MA: Harvard University Press, 1998.

[12] Schroeder, M., "Reasons First," Oxford University Press, 2021.

[13] Tucker, C., *A holist balance scale*, Journal of the American Philosophical Association **First View** (2022), pp. 1–21.

[4] For the latter, see, e.g., Dietrich and List [2], Faroldi and Protopescu [3], and Horty [4].

[5] It pays noting that our approach is similar to the methodology underlying input/output logic [6,7] which is built around factual detachment—see [8, pp. 502–5] for a discussion here.

Autonomous decisions and rational choice: How to model epistemic rationality?

Nicolas Tardif [1]

Université du Québec à Trois-Rivières
Trois-Rivières, QC (Canada)

Clayton Peterson [2]

Université du Québec à Trois-Rivières
Trois-Rivières, QC (Canada)

Keywords: Ethical AI, Machine ethics, Automated decision procedures, Ethical agents, Rational Choice

Many scholars think that artificial ethical (moral) agents (otherwise known as AMA), capable of autonomous ethical decision making, can be defined (cf. [2],[12], [14], [15]). In the machine ethics literature, rational choice theory, specifically understood on the grounds of instrumental rationality, is generally conceived as a prerequisite for these artificial ethical agents (cf. [13], [9], [18], [1]), which are (at least partly) conceived as expected utility maximizers. Peterson [16] recently showed some limitations of implementing rational choice through the maximization of expected utility. While instrumental rationality is generally conceived as the only type of rationality needed to implement in autonomous decision making in the machine ethics literature, Peterson showed how parameters and coding affect the output of the alleged ethical choice, incidentally pointing out the insufficiency of instrumental rationality for modeling ethical choices. As such, assuming that artificial ethical agents can be defined (which is in itself arguable, see [17]), instrumental rationality might not be the only type of rationality that needs to be incorporated within models of autonomous reasoning. For instance, the epistemic norms regulating how one should obtain, revise, or discard beliefs also plays a significant role in rational choice. While some scholars argue that epistemic rationality can be reduced to instrumental rationality, the relationship between these two types of rationality

[1] M.A. candidate, Nicolas.Tardif2@uqtr.ca

[2] This work was financially supported by the *Fonds de Recherche du Québec* and the *Fonds National de Recherche du Luxembourg* [FRQSC-AUDACE-310754] as well as by the *Fonds de Recherche du Québec* [2023-NP-310505].

is actually not trivial (see for example [3], [5], [4], [6], [7], [8], [10], [11], [19], [20], [21]). In light of these considerations, our objective is to get a deeper understanding of the relationship between instrumental and epistemic rationality and determine the limitations of these types of rationality in modeling autonomous ethical decision making.

References

[1] Abel, D., J. MacGlashan and M. L. Littman, *Reinforcement learning as a framework for ethical decision making*, in: *AAAI Workshop: AI, Ethics, and Society*, 2016 .

[2] Anderson, M. and S. L. Anderson, *Machine ethics: Creating an ethical intelligent agent*, AI magazine **28** (2007), pp. 15–15.

[3] Berker, S., *Epistemic teleology and the separateness of propositions*, The Philosophical Review **122** (2013), p. 337–393.

[4] Christensen, D., *The ineliminability of epistemic rationality*, Philosophy and Phenomenological Research **103** (2021), pp. 501–517.

[5] Cowie, C., *In defence of instrumentalism about epistemic normativity*, Synthese, **191** (2014), pp. 4003–4017.

[6] Kelly, T., *The rationality of belief and some other propositional attitudes*, Philosophical Studies **110** (2002), pp. 163–196.

[7] Kelly, T., *Epistemic rationality as instrumental rationality: A critique*, Philosophy and Phenomenological Research **66** (2003), pp. 612–640.

[8] Kelly, T., *Evidence and normativity: Reply to leite*, Philosophy and Phenomenological Research **75** (2007), pp. 465–474.

[9] Kochenderfer, M. J., "Decision Making Under Uncertainty: Theory and Application," MIT Press, 2015.

[10] Leite, A., *Epistemic instrumentalism and reasons for belief: A reply to tom kelly's "epistemic rationality as instrumental rationality: A critique"*, Philosophy and Phenomenological Research **75** (2007), pp. 456–464.

[11] Lockard, M., *Epistemic instrumentalism*, Synthese **190** (2013), p. 1701–1718.

[12] Misselhorn, C., *Artificial moral agents: Conceptual issues and ethical controversy*, in: S. Voeneky, P. Kellmeyer, O. Mueller and W. Burgard, editors, *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*, Cambridge University Press, 2022 p. 31–49.

[13] Moor, J. H., *Just consequentialism and computing*, Ethics and Information Technology **1** (1999), pp. 61–65.

[14] Moor, J. H., *The nature, importance, and difficulty of machine ethics*, IEEE Intelligent Systems **21** (2006), pp. 18–21.

[15] Muehlhauser, L. and L. Helm, *The singularity and machine ethics*, in: A. Eden, J. Moor, J. Søraker and E. Steinhart, editors, *Singularity Hypotheses*, The Frontiers Collection, Springer, 2012 pp. 101–126.

[16] Peterson, C., *Further thoughts on defining $f(x)$ for ethical machines: Ethics, rationality, and risk analysis*, The International FLAIRS Conference Proceedings **36** (2023).

[17] Peterson, C. and N. Hamrouni, *Preliminary thoughts on defining $f(x)$ for ethical machines*, The International FLAIRS Conference Proceedings **35** (2022).

[18] Russell, S. and P. Norvig, "Artificial Intelligence: A Modern Approach," Global Edition, 2022, 4th edition.

[19] Siegel, H., *Epistemic rationality: Not (just) instrumental*, Metaphilosophy **50** (2019), pp. 608–630.

[20] Skipper, M., *Unifying epistemic and practical rationality*, Mind **132** (2023), pp. 136–157.

[21] Vahid, H., *Rationalizing beliefs: evidential vs. pragmatic reasons*, Synthese **447-462** (2010), p. 3.

Rule rather than Exception: Defeasible Probabilistic Dyadic Deontic Logic.

Vincent de Wit [1]

University of Luxembourg
Department of Computer Science
Esch-sur-Alzette, Luxembourg

Keywords: Deontic Logic, Probabilistic Logic, Multi-Agent, Formal Argumentation, Theory of Mind.

We take probabilistic deontic logic [2] and make it defeasible using an argumentation method published in [3]. The specific probabilistic deontic logic we use is dyadic deontic logic [6][5][9] combined with a multi-agent variant of probabilistic logic [4]. More specifically, we will use the developed method of specifying an upper and a lower bound logic that will define the strict and defeasible rules of an $ASPIC^+$ framework [8][1]. The lower bound logic uses axiom system G of the Hansson-Lewis systems of Dyadic Deontic Logic combined with axioms for the probabilistic logic. And for the upper bound logic, the axioms of Upward and Downward inheritance introduced in [10] are added.

We consider the described framework as a framework that is used by an agent to describe specific elements of its surroundings and reason about it. The framework combines multiple operators namely: strict and defeasible implications ($\rightarrow$, $\Rightarrow$), permissions and obligations ($O(\phi|\psi)$, $P(\phi|\psi)$), agent specific probabilistic formulas ($\alpha * w_i(\phi) \geq \beta$), theory of mind formulas ($\alpha_1 * w_i(\alpha_2 * w_j(\phi) \geq \beta_2) \geq \beta_1$) and also strict and defeasible knowledge. An important question therefore is: “What is the difference between a defeasible permission and an uncertain permission?” The difference is that defeasible knowledge is debunkable while the uncertainty about something is not debunkable, if the uncertainty is not defeasible in the first place -Normally ϕ is permitted; I am uncertain whether ϕ is permitted–. Furthermore, we

[1] e-mail: vincent.dewit@uni.lu

[2] Vincent de Wit is supported by the Luxembourg National Research Fund PRIDE19/14268506 .

[3] This work was supported by the Fonds National de la Recherche Luxembourg through the project INTEGRAUTO (INTER/AUDACE/21/16695098).

can express “normally I am uncertain whether ϕ is permitted.” The work also opens up whether the framework should give preference to more certain information, i.e. $w_i(\phi) \geq 0.8$ versus $w_i(\psi) \geq 0.7$. There is more justification for ϕ than there is for ψ.

Consider the following example scenario of an agent learning about the rules of a library. The agent does not know that it is a rule to be silent in a library, and attempts to derive such rules without breaking them or explicitly asking other agents about the rules. While in the library, the agent will discover that most people are silent in the library, though that there are exceptions –for example at the checkout counter–. Furthermore, the agent will encounter an ambiguous situation in which people talk inside a room. Multiple explanations are possible in this case: it is allowed to talk inside the room, the people do not know about the rule to be silent, or the people do not care about the rule i.e. they are breaking the rule.

Furthermore, we set out to determine whether this framework is able to solve the paradox of epistemic obligation satisfyingly. A paradox of epistemic obligation goes as follows: (1) The bank is being robbed; (2) It ought to be the case that Jones (the guard) knows that the bank is being robbed; (3) It ought to be the case that the bank is being robbed [7]. The proposed solution is to consider the formulas in the knowledge base as known by the agent.

Lastly, notable future research is whether it is possible to lift syntax level probability to probability on argument level.

References

[1] Baroni, P., D. Gabbay, M. Giacomin and L. v. d. Torre, “Handbook of Formal Argumentation,” London, England: College Publications, 2018.

[2] de Wit, V., D. Doder and J. Meyer, **12897 LNAI**, 2021.

[3] Dong, H., B. Liao, R. Markovich and L. van der Torre, *Defeasible Deontic Logic: Arguing about Permission and Obligation*, Journal of Applied Logics **9** (2022).
URL `https://orbilu.uni.lu/handle/10993/54193`

[4] Fagin, R. and J. Y. Halpern, *Reasoning about knowledge and probability*, J. ACM **41** (1994), pp. 340–367.
URL `https://dl.acm.org/doi/10.1145/174652.174658`

[5] Gabbay, D., J. Horty, X. Parent, R. van der Meyden and L. van der Torre, *Handbook of deontic logic and normative systems* (2013).

[6] Hansson, B., *An Analysis of some Deontic Logics*, Noûs **3** (1969), pp. 373–398.
URL `https://www.jstor.org/stable/2214372`

[7] Hulstijn, J., *Need to know: Questions and the paradox of epistemic obligation*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **5076 LNAI** (2008), pp. 125–139.
URL `https://link.springer.com/chapter/10.1007/978-3-540-70525-3_11`

[8] Modgil, S. and H. Prakken, *The ASPIC + framework for structured argumentation: a tutorial*, Argument & Computation **5** (2014), pp. 31–62.
URL `https://content.iospress.com/articles/argument-and-computation/869766`

[9] Parent, X. and L. van der Torre, "Introduction to Deontic Logic and Normative Systems," College Publications, 2018.
URL `https://orbilu.uni.lu/handle/10993/40374`

[10] Prakken, H. and M. Sergot, *Contrary-to-duty obligations*, Stud Logica **57** (1996), pp. 91–115.
URL `http://link.springer.com/10.1007/BF00370671`