

The Use of Agent-based Simulation of Public Policy Design to Study the Value Alignment Problem

Pablo Noriega¹, Enric Plaza¹

¹IIIA-CSIC, Barcelona, Catalonia, Spain

Abstract

We propose to use agent-based simulation (ABS) of public policies to explore fundamental and practical issues associated with the role of values in the governance of autonomous artificial systems.

Keywords

Value Alignment Problem (VAP), AI governance, value engineering, agent-based simulation, public policy design


1. Introduction

The raison d'être of Artificial Intelligence is the design and construction of autonomous artefacts. Such autonomy is the source of AI's main contributions and concerns, hence the relevance of devising ways to harness it. One way to achieve this is to engineer values into Artificial Intelligent Systems. More technically, to address the "Value Alignment Problem" (VAP), that S. Russell characterised as "designing and building autonomous artificially intelligent systems (AIS) whose behaviour is, "provably aligned with human values" [1].¹ This paper is an argument in favour of an experimental approach to VAP, using agent-based simulation (ABS) of public policy design (Sec. 2) to identify the the key components of VAP (Sec. 3) and explore the way value engineering can be addressed in practice (see Sec. 4). We have discussed some of these ideas before in [3, 4] and Perelló's PhD. dissertation [5] develops some of the underlying intuitions. We illustrate our proposal with the case of residential use of water in a medium size city involving households that demand and use water, a utility company that provides the service and the city government that oversees that the service is provided according to policy.²

Proceedings of Artificial Intelligence Governance Ethics and Law (AIGEL), Reviewed, Selected Papers. November 02 - December 19, 2022, Barcelona, Spain

✉ pablo@iiia.csic.es (P. Noriega); enric@iiia.csic.es (E. Plaza)

🆔 0000-0003-1317-2541 (P. Noriega); 0000-0003-1283-8188 (E. Plaza)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹We find that Russell's phrasing of VAP's in terms of alignment with human values (in [1]) is better suited for technical discussion than the more generic beneficial (in [2]). Moreover, we understand "provably aligned" not in proof-theoretic terms but rather as an objective way of measuring to what extent the behaviour of an AIS is consistent with a specif set of values.

²See [6, 5] for a detailed discussion

2. Modelling policy design as a value-alignment problem

Given a policy domain (urban use of water) and a group of stakeholders (households, utility company and city hall), a policy is an intervention that intends to improve the state of affairs. Thus, policy design involves the identification and the articulation of means and ends (that conform a policy intervention), followed by an assessment that such intervention is actually conducive to the intended improvement [7].

Values determine, in the policy itself, what is an improvement, whether an intervention succeeds in achieving the improvement through appropriate instruments, and whether stakeholders are satisfied with the intervention. Consequently, policy design has to model, on the one hand, how values are involved in the decision-making of those individual agents whose collective activity is being affected by a policy; and, on the other, how values are involved in the governance of that collective behaviour —that is, in the design of the policy itself.

Policy design is a complex problem —mainly because several variables with complex interactions determine the activity and its effects and several (often conflicting) motivations and interests are in play— involving factual and ethical decisions (cf. H. Simon [8]). Simulation, being able to deal with such complexity and trade-offs experimentally, is arguably a reasonable methodological approach to policy design [9, 4]. Agent-based simulation (ABS) is an appropriate type of simulation for policy design because it separates design concerns in the modelling of individuals and in the modelling of collective action.

In this context, ABS for policy interventions can be seen as a particular form of the VAP: It is a design process problem with the two main tasks embedding values in an AIS and assessing that the behaviour of the system is objectively aligned with those values. In fact, however, value-driven policy modelling has the added advantage of a dual perspective of value embedding: how to embed values in the decision model of an agent; and how to embed values in the means and ends of a policy intervention. Alignment, in turn, can be studied in the outcomes of the simulation by looking at the degree of satisfaction of individual agents with the outcomes of the policy with respect to the agent's individual values, and in the effectiveness of the policy intervention in the fulfilment of the postulated values. In fact, ABS provides an experimental framework for testing the adequacy of these models.

3. Making the VAP operational in policy modelling

From the perspective of policy design, the point of ABS is to identify a course of action that is effective in reaching some desirable goals, through means that have reasonable trade-offs and are acceptable to stakeholders. Values play a key role in this process because they provide the central elements of the simulation. In fact, having some explicit values in mind, suggest which elements must be observable through the simulation (in order to assess and compare outcomes of simulation runs). Values also serve to identify what actions are conducive to desirable or undesirable outcomes, and therefore values determine not only that such actions be included in the simulation but also to include the governance means that harness policy subject actions toward the desirable outcomes. Along these lines, values should be involved in the modelling of policy subjects: an agent chooses to perform an action when that the action satisfies the needs

and preferences (i.e. values) of that individual. Finally, as suggested in the previous section, values determine different perspectives for the assessment of the outcomes of simulation runs; thus, one needs to model whether an intervention leads to desirable states-of-the-world, the relative advantages of (equally effective) interventions, and whether a given intervention is more or less compatible with the values and preferences of the different stakeholders (including direct policy-design stakeholders as well as individual policy subjects).

In order to make this role of values operational for simulation purposes one needs to address two design concerns: on the one hand, modelling of the policy domain and the behaviour of agents; and on the other, engineering of values into those models (Fig. 4). In technical terms the policy domain is modelled as an online institution and policy subjects as the autonomous agents that interact within it (see [10]). The engineering of values can be addressed as outlined below and discussed in [11].

Modelling policy domain and policy. subjects Without going into details, the policy domain needs to include a Physical Model –that captures what happens in that fragment of the real world (the policy domain) that has to be taken into account for the design of a policy– and a Governance Model –that captures the artificial constraints that govern the interactions of policy subjects. For example, in the case of urban water policy, the Physical Model contains an abstraction of the real-world conditions that correspond to the supply and use of water in a city (the total supply of water, a number of households with their specific economic, demographic and water use profiles, the cost of treating a ton of water, how many litres are used to take a regular shower, how much water is saved with an ecological toilet and the cost of buying one; and so on). The Governance Model includes the conventions (norms, regulations and even social practices) that bear upon water use and supply (the way invoices are calculated and presented to households, the standards for water quality, subsidies for refurbishing household appliances, contracting conditions for water utilities, etc).

The simulation of policy subjects usually amounts to some assumptions about the population of agents and the modelling of their decision-making. The core modelling assumes –in our proposal– that an agent takes a policy-enabled action only when opportunity, capability and motivation concur. In our example, households are defined by a socio-demographic profile (data like the number of household-members, age, sex, income; based on empirical data like census, administrative actions, etc) and a value profile (that characterises the preferences and priorities of households; based on more or less standard value taxonomies like Schwartz [12]) that will be involved in modelling the motivation part of decision-making.³

Policy intervention. A policy intervention is a selection of policy instruments whose effects on the state of the world are to be assessed (e.g., introducing an incrementally progressive fare for household water use together with subsidies for purchasing water-saving devices and a campaign to foster the adoption of water-saving practices in order to reduce household water consumption).

Engineering values. The point of this process is to translate an abstract notion of value into concrete constructs that may be embedded in a policy intervention, in the domain model

³The modelling of capability is linked to the socio-demographic profile of the agent and constrained by the governance conventions. Opportunity has to do with the physical constraints, the state of the world and the current conditions of the household profile variables.

and as part of the agent decision-model of individual policy subjects.

A. The translation process. It can be organised as a cycle with three main stages: value choice, value interpretation and value-alignment assessment, as follows:

1. Value choice. Identifying those values that are appropriate for the policy domain and those stakeholders whose values ought to be represented in the policy that is being designed. Urban water use would prioritise values like sustainability, healthiness, security, fairness, efficiency, etc; and the stakeholders are not only the policy subjects that we are explicitly modelling (households and water utility companies) but also those that are involved in the policy design (the city administration that will be responsible for the policy deployment and follow up, the politicians who promote and negotiate the policy and other indirect stakeholders like industry and agriculture, climate advocacy groups, banks).

2. Value interpretation. This stage addresses two problems:

1. *Making values observable.* That is, turn a label that stands for an abstract value into a feature that is measurable. Namely a goal or a set of goals that is motivated and legitimised by the value. It is convenient to distinguish goals that are consensual (because they correspond to values that should be embedded in the policy, independently of the individual values of stakeholders) and those goals that are desirable for each stakeholder. For instance, in our example, the goal to reduce individual water consumption to a certain per-capital-annual volume may be one goal that stands for the consensual value of sustainability.
2. *Instrumenting values.* That is, identify means to achieve the intended goals. Since the actions of agents is what leads or detracts from the satisfaction of goals, the instrumentation of a value is the modulation of those actions that affect those goals. Thus to achieve sustainability, one may want to promote the adoption of water saving devices as a way of reducing individual water use; and for this purpose a city may decide to regulate sanitation standards for new housing or provide subsidies for retrofitting and start a campaign to motivate such adoption. In fact, we assume that instruments can only be of three types: afford or prevent actions (specified in the physical model); promote or discourage actions (in the governance model), and provide information that may facilitate the decision-making of participants towards those actions (from the governance model and through the physical model).

Figure 1 shows the goal decomposition process that starts with the consensual policy-maker's domain-specific values (on the left) until reaching (on the right) the salient policy goals. For each policy goal, the diagram also shows some of the corresponding observable indicators and some instruments that impact on those indicators.

3. Value-alignment assessment. Establish the conventions to determine to what degree a specific value is being fulfilled and then define a way of combining the satisfaction of a set of values to determine the degree to which all the values are being fulfilled. This can be achieved in different ways but one may think of this alignment assessment as a multi-objective optimisation process of sorts. First, for each value one can postulate a threshold of satisfaction (that we call an "aspiration level") —e.g. a 15% reduction of domestic water use over the next five years— and a way of qualifying to what extent any potential state of affairs may be better or worse

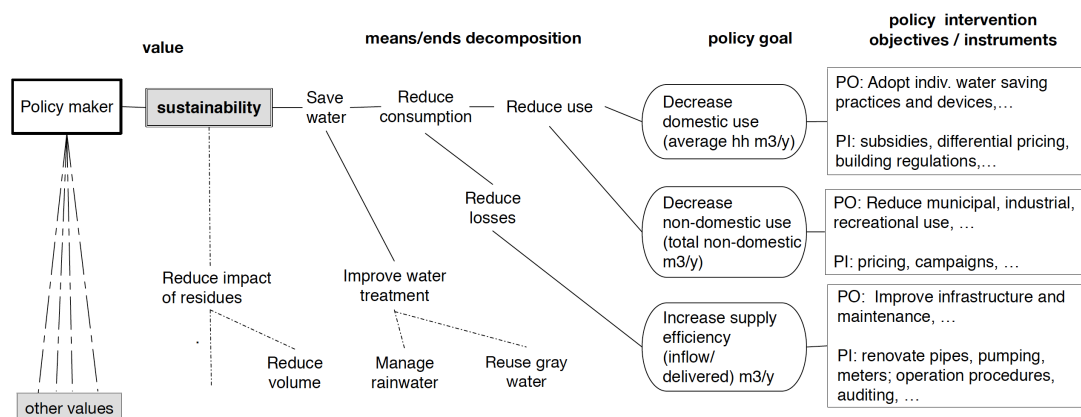


Figure 1: From values to policy goals and instruments. The diagram shows the top part of a graph of consensual value interpretation for the design of an urban water use policy.

—a utility-like function, for example, that ranks different water reduction rates— thus defining some kind satisfaction function for each value. Second, one needs an aggregation function that determines the degree of satisfaction of all the values —for instance the satisficing aggregation function in Fig. 2.

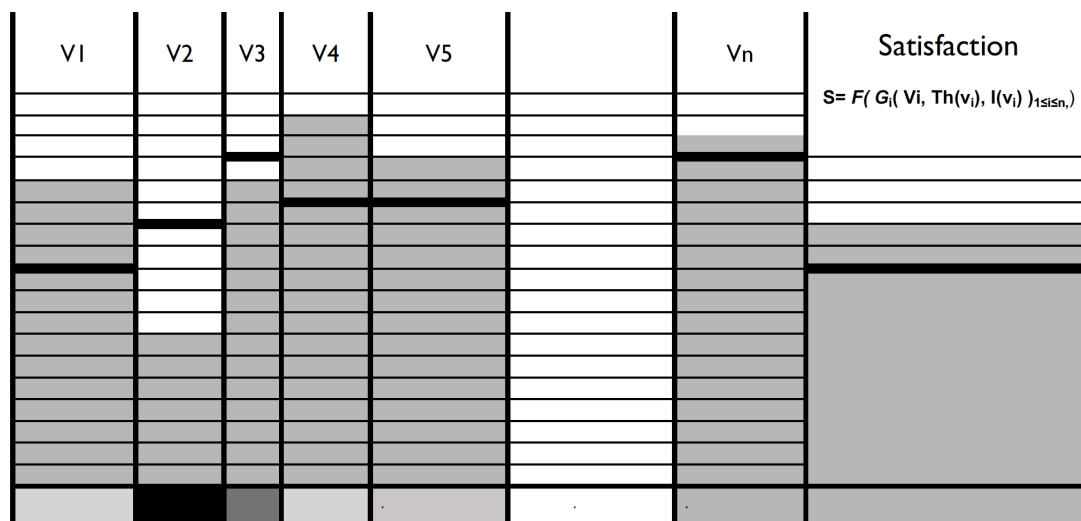


Figure 2: A simple value aggregation satisficing function. Each value (v_1, \dots, v_n) has a threshold, or aspiration level (dark line), a relative importance with respect to other values (width of the column) a relative fulfilment in a given state of the world (coloured cells): Their aggregated satisfaction is represented in the last column. In this case, the state of the world is “effectively aligned” with the set of values since the satisfaction with most values compensate the dissatisfaction with values V_2 and V_3 of lesser relative importance.

We propose three kinds of assessment of a policy intervention: effectiveness (to what extent the values are being served), adequacy (trade-offs in the choice of instruments), and acceptability

(alignment with stakeholders' values) —see [13] for a similar distinction of relevant assessment criteria, specific to the water domain.

B. Embedding values in the models. Once the values are translated into concrete constructs through the process just described, these are made operational in two contexts: First, in the modelling individual value-aligned behaviour in each simulated policy subject. This can be accomplished in three ways: as a deterministic reaction to certain situations, as a learning mechanism that leads to consistent behaviour that is value aligned, or as a cognitive component in the agent reasoning that makes values bear upon the selection of actions, as illustrated in Fig. 3.

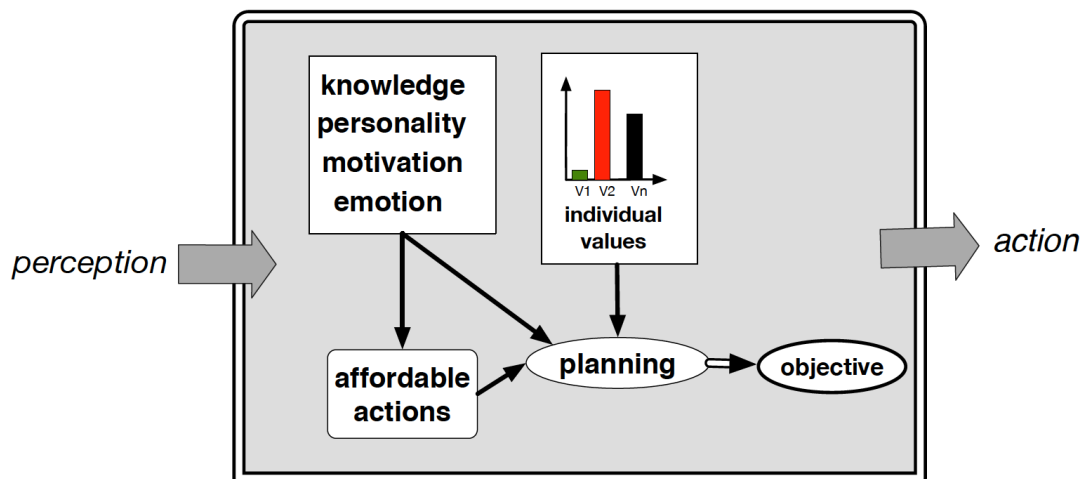


Figure 3: Value-driven decision model for an individual household. A household attempts an afforded action only when it has the opportunity, the capability and the motivation to attempt it. Values are part of the mechanism that models motivation.

Second, values become operational in the policy itself: as part of the policy domain model —as affordances and constraints of the physical and governance models— as part of a policy intervention (through the features that serve to measure values and the selection of policy instruments), and as the way of assessing the alignment of an intervention (as sketched in Fig. 4).

4. Agent-based simulation of policy interventions

The purpose of using ABS to design a policy is to provide experimental evidence to support the choice of a policy intervention. The point is to test, in a systematic way, different policy interventions —combination of policy instruments— and identify the ones that lead to the best end results. Fig. 4 summarises the modelling assumptions and describes the simulation cycle.

Experimental settings. There are four main components:

1. *Starting conditions and evolution.* These include:

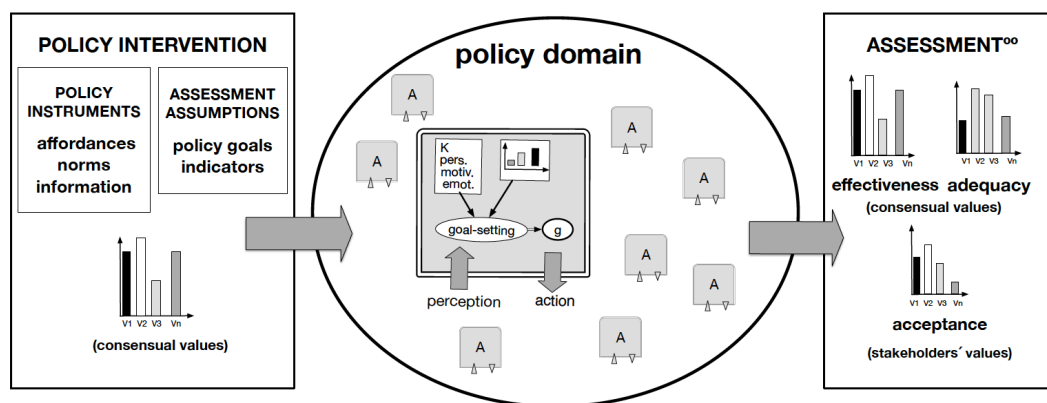


Figure 4: The role of values in policy modelling. Values are embedded in the policy domain model (in the form of constraints and affordances) and are reflected in the way value alignment is assessed and in the specification of a policy intervention.

- (i) all the parameters of the physical model that hold at time $\exists 0$ (including a starting population of policy subjects);
 - (ii) all the governance conventions (norms, regulations, enforcement mechanisms) that are active at time $\exists 0$ and their corresponding parameters; and
 - (iii) a specification to account for the evolution of the state of the world (e.g. environmental, economic and demographic changes).
2. *Policy intervention* (or, more intuitively, a sensible list of instruments). For each instrument in the intervention:
 - (i) the list of indicators of the state of the world that are affected with the enactment of the instrument;
 - (ii) the parameters associated with each instrument that are worth testing; and
 - (iii) a measure of the most significant positive and negative effects of the instrument on indicators and policy goals.
 3. *Assessment assumptions*. We distinguish assessment with respect to the consensual values and with respect to the values of the individual stakeholders.
 - (i) *Consensual value interpretation*: The definition of (consensual) policy goals and their assessment features: aspiration levels, degree of satisfaction and aggregation functions. These assumptions are used in the effectiveness and adequacy assessments
 - (ii) *Value interpretations of stakeholders*: The goals, satisfaction and aggregation functions that are specific for each of those stakeholders who participate directly in the design of the policy (city hall, utility companies, special interest groups, and so on). These assessment features are involved in the adequacy and acceptability assessments.

- (iii) Policy subjects' value interpretation: The values and assessment features that belong to each simulated policy subject type. These will be used in their individual decision models and in the acceptability assessment.
4. *Assessment functions.* We will use three measures to assess different aspects of achieving a goal.
- (i) Effectiveness, to measure how successful is the policy intervention in achieving the consensual goals;
 - (ii) Adequacy, that determines whether the means to achieve those goals are adequate in terms of their collateral effects (in order to assess the trade-offs between interventions);and
 - (iii) Acceptance, that measures the degree to which the policy intervention aligns with the values of simulated policy subjects as well as with those of the direct stakeholders that participate in the policy design and negotiation process.

Testing cycle. Given a set of starting conditions a policy intervention is evaluated with respect to a set of assumptions about value interpretation, and assessment. Simulation allows to explore the effects of changing parameters of the policy intervention instruments, changing the instruments, changing value choices, interpretation, instrumentation and assessment, and also modifying the starting conditions. These experiments are meant to provide support for the comparison of policy interventions, and in this way contribute evidence towards policy negotiation and deployment.

5. Closing remarks

1. Values as an explicit design feature. As we argued in the introduction values are an essential feature of policy design. As far as the modelling process is concerned, values elucidate what entities (objects) need to become part of the physical model of the domain, what entities need to be observable in the simulated state of the world, what actions need to be afforded and what constraints need to be implemented. Values may also be used to elucidate the analogous features in the design of a large variety of artificial intelligent systems. For instance, the ideas we outlined in this paper are directly applicable to artificial intelligent systems that involve the online coordination of autonomous agents, as discussed in [11].

2. An experimental approach to value engineering. In the previous section, we gave a shallow description of the process of engineering values into a simulated multiagent system. This outline would need to be fleshed out for designing a particular policy through sound simulation. However, these ideas may also be expanded to support an experimental approach to value engineering by exploring alternative ways to address each of the tasks in the valueengineering process. For instance alternative value aggregation functions; agent architectures for ethical reasoning and so on. Such exercises should provide evidence for design guidelines and heuristics for value engineering.

3. Value-based governance of artificial systems. This paper can be read as an exercise of modelling a value-based policy design process. However, the way we chose to model the policy domain and the autonomous policy users can be extended to modelling of online systems that involve autonomous agents that may be artificial or not. This way, the affordances and constraints as well as the instruments that guide a policy intervention can be understood as value-driven governance means that harness the autonomous behaviour of those entities within the online system. Likewise, the way values are embedded in the decision-making of policy subjects in this paper is but an simplified example of the process of engineering value-driven behaviour into an autonomous artificial agent.

4. An AI-inspired theory of values. This paper is also an argument in favour of the exploration of an AI-inspired theory of values. In particular, we envisage a re-examination of conventional views on values in such a way that values may also be ascribed to artefacts having self-driven behaviour within a social context. We claim one needs to take into account four core concepts to articulate such theory: values, autonomy, governance, and collective action. The interplay among these four concepts shapes the research landscape in which AI systems can be value abiding. From a methodological standpoint, we believe that the approach to the development of an AI-inspired theory of values may profit from available AI developments and mirror the path followed in classical AI: a few well-chosen core concepts, multidisciplinary work, building science along with engineering, and designing paradigmatic problems –and, as we argue in this paper, value-driven policy design may be one of these.

Acknowledgments

Research for his paper is supported by EU (HORIZON-EIC-2021-PATHFINDERCHALLENGES-01) Project VALAWAI 101070930; the EU (NextGenerationEU/ PRTR program) and the Spanish (MCIN/AEI /10.13039/501100011033 program) project VAE TED2021-131295B-C31; and CSIC's (Bilateral Collaboration Initiative i-LINK-TEC) project DESAFIA2030 BILTC22005.

References

- [1] S. Russell, Of Myths and Moonshine. A conversation with Jaron Lanier, 14-11-14, The Edge, 2014. URL: <https://www.edge.org/conversation/the-myth-of-ai#26015>, [Online] Retrieved 12 december 2022.
- [2] S. Russell, Provably beneficial artificial intelligence, *The Next Step: Exponential Life*, BBVA-Open Mind (2017).
- [3] A. Perello-Moragues, P. Noriega, A playground for the value alignment problem, in: L. Martínez-Villaseñor, I. Batyrshin, A. Marín-Hernández (Eds.), *Advances in Soft Computing*. Mexican International Conference on Artificial Intelligence, volume 11835 of LNCS, Springer, 2019, pp. 414–429. doi:https://doi.org/10.1007/978-3-030-33749-0_33.
- [4] A. Perello-Moragues, P. Noriega, Using agent-based simulation to understand the role of values in policy-making, in: *Advances in Social Simulation*, Springer, 2020, pp. 355–369. doi:https://doi.org/10.1007/978-3-030-34127-5_35.

- [5] A. Perello-Moragues, P. Noriega, A. Popartan, M. Poch, Modelling policy shift advocacy, in: Proceedings of the Multi-Agent-Based Simulation Workshop (MABS) in the International Conference on Autonomous Agents and Multiagent Systems (AAMAS19), In Press.
- [6] A. Perello-Moragues, M. Poch, D. Sauri, L. A. Popartan, P. Noriega, Modelling domestic water use in metropolitan areas using socio-cognitive agents, *Water* 13 (2021). URL: <https://www.mdpi.com/2073-4441/13/8/1024>. doi:10.3390/w13081024.
- [7] P. J. May, Policy design and implementation, in: B. Peters, J. Pierre (Eds.), *The SAGE Handbook of Public Administration*, 2nd ed., SAGE Publications, 2012, pp. 279–291.
- [8] H. A. Simon, Fact and Value in Decision-making, in: *Administrative Behavior: A study of decision-making processes in administrative organization*, 4th ed., The Free Press, 1997.
- [9] N. Gilbert, P. Ahrweiler, P. Barbrook-Johnson, K. P. Narasimhan, H. Wilkinson, Computational modelling of public policy: Reflections on practice, *Journal of Artificial Societies and Social Simulation* 21 (2018) 14.
- [10] P. Noriega, J. Padget, H. Verhagen, M. d’Inverno, Anchoring online institutions, in: P. Casanovas, J. J. Moreso (Eds.), *Anchoring Institutions. Democracy and Regulations in a Global and Semi-automated World*, Springer, (in press).
- [11] P. Noriega, H. Verhagen, J. Padget, M. d’Inverno, Design heuristics for ethical online institutions, in: N. Ajmeri, A. Morris Martin, B. T. R. Savarimuthu (Eds.), *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV*, Springer International Publishing, Cham, 2022, pp. 213–230.
- [12] S. H. Schwartz, Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries, in: *Advances in experimental social psychology*, volume 25, Elsevier, 1992, pp. 1–65.
- [13] C. Perry, ABCDE+F: A framework for thinking about water resources management, *Water International* 38 (2013) 95–107.