# A Playground for the Value Alignment Problem

Antoni Perello-Moragues[1,2,3(✉)] and Pablo Noriega[1]

[1] IIIA-CSIC, Barcelona, Spain
{tperello,pablo}@iiia.csic.es
[2] Aqualia, Barcelona, Spain
antonio.perello@fcc.es
[3] Universitat Autònoma de Barcelona, Barcelona, Spain

**Abstract.** The popularity of some recent applications of AI has given rise to some concerns in society about the risks of AI. One response to these concerns has been the orientation of scientific and technological efforts towards a responsible development of AI. This stance has been articulated from different perspectives. One is to focus on the risks associated with the autonomy of artificial entities, and one way of making this focus operational is the "value-alignment problem" (VAP); namely, to study how this autonomy may become provably aligned with certain moral values. With this purpose in mind, we advocate the characterisation of a problem archetype to study how values may be imbued in autonomous artificially intelligent entities. The motivation is twofold, on one hand to decompose a complex problem to study simpler elements and, on the other, the successful precedents of this artifice in analogous contexts (e.g. chess for cognitive AI, RoboCup for intelligent robotics). We propose to use agent-based modelling of policy-making for this purpose because policy-making (i) constitutes a problem domain that is rich, accessible and evocative, (ii) one may claim that it is an essentially value-drive process and, (iii) it allows for a crisp differentiation of two complementary views of VAP: imbuing values in agents and imbuing values in the social system in order to foster value-aligned behaviour of the agents that act within the system. In this paper we elaborate the former argument, propose a characterisation of the archetype and identify research lines that may be systematically studied with this archetype.

**Keywords:** Ethics in AI · Agent-based simulation · Policy-making · Policy values

## 1 Introduction

There is concern in society about the dangers of AI [6,44]. Although concerns may be founded, we propose to see beyond the Frankenstein/Terminator image and focus on dangers that are specific to AI and how to address them. This more


L. Martínez-Villaseñor et al. (Eds.): MICAI 2019, LNAI 11835, pp. 414–429, 2019.
https://doi.org/10.1007/978-3-030-33749-0_33

focused view has been adopted by a large part of the AI community and one may recognise three main strategies to address these concerns:

**Strategy 1.** It consists in taking a classical moral approach that does, in the field of AI, what is usually done in other fields: emphasis in a proper ethical education and developing a sense of responsibility that is specific to the field. This strategy is reflected in several manifestos and put into practice through institutions, programs and projects that promote an AI that has been qualified as "human-centred", "responsible" or "ethical". For instance, [12,17,34].

**Strategy 2.** With a crisp understanding of AI potential, it directs efforts towards the formulation, promotion and adoption of legislation, guidelines, principles, standards, certification mechanisms and similar ways of identifying specific risks that may be better characterised, allocated prevented and eventually minimised or repaired. For instance [22].

**Strategy 3.** It adopts the stance of taking advantage of the power of AI to solve the problems caused by AI. This strategy is following several paths; one is to "teach computers to do the right thing" [3,6]. A second articulation is to find ways to "guarantee" that machines behave morally or to build autonomous systems that are "provably aligned" with a set of values [36]. This last formulation is what is called the *value alignment problem* (VAP). Although it is unlikely that the value alignment problem may be solved for every autonomous system, it is plausible to characterise a class of autonomous systems that may be imbued with values and to develop formal and methodological means to eventually support their provability.

With this aim in mind, we advocate a restricted version of VAP that facilitates addressing several of these underlying issues. The point is to establish an objective frame of reference where ethical questions may be worked out and eventually produce insights that may apply to value aligned behaviour within some well defined conditions. We propose to use the problem of policy-making and to explore this problem through agent-based simulation. The choice of subject matter is based on the understanding that policy-making is a value-driven activity, that involves two complementary perspective of values: first, the social values that a policy-maker seeks to imbue in a social system and, second, the values imbued in the behaviour of individual policy-subjects. The choice of agent-based modelling, as we argue in Sect. 2, is a methodological decision that fosters a bare-bones description of the problem that may be appropriate for a crisp formulation of some questions and an objective assessment of possible answers.

The approach of characterising a "landmark" model for complex scientific problems has proven effective before in several cases. In the case of AI, chess-playing for symbolic cognitive AI and *Robocup* for intelligent robotics come immediately to mind. Our proposal is more modest in scope, but the underlying questions are no less significant. Thus, the questions that may be addressed with this landmark model should promote several research and development lines and that the impact of the research and development spawned by it should have a beneficial impact on the development of an ethically responsible AI.

This paper is essentially an elaboration of these matters and hence it is structured as follows. In Sect. 2 we advocate the suitability of agent-based modelling (ABM) for value-driven policy-making. In Sect. 3 we outline a conceptual framework for value-driven policy-making. In fact, we propose a "metamodel" that establishes the components needed to characterise the archetype problem, we then propose a refinement of that metamodel to characterise a more concrete form of the archetype that we call a *landmark class* and, in order to end up with restricted enough versions of the archetype we choose a particular landmark subclass to define a *challenge*. Finally, in Sects. 4 and 5 we outline a research programme around the type of questions for which the archetype, or more specifically its challenge subclass, provide motivation and the grounds to develop theory and technology to imbue values in autonomous intelligent systems. A summary of the background of our proposal is in Appendix.

## 2 Why ABM of Value-Driven Policy-Making?

Policy-making is, intrinsically, a value-driven process: policy-makers design policies in order to achieve a "better—with respect to some value system—state of the world". In order to achieve that goal, policy-makers make use of means of different sorts that foster behaviour of other stakeholders towards those ends. Policy stakeholders decide what are the "right" actions to take and assess the "goodness" of the state of the world but with respect to their own values.

The design of a public policy separates value alignment in two complementary perspectives: the problem of imbuing values in the global emergent behaviour of systems involving autonomous entities, on one hand; and the problem of imbuing values in the behaviour of individual autonomous agents. In the first case, policy instruments foster value aligned behaviour and policy objectives make explicit those values that intend to be achieved. In the second case, one may study how values are involved in the decision making of individuals and how their behaviour is aligned to their own individual values and to the values fostered by the social policy. Each perspective allows for exploration of different questions. By choosing a particular policy domain many conceptual and methodological concerns become easier to handle and one may get an empirical grounding and validation for some assumptions.

Agent-based modelling is a convenient way to explore these matters because it provides a refutational framing of value-related modelling. In fact, it motivates identifying and objectifying value-related assumptions, it provides a test-bed for operationalisation of core value notions (value assessment, representation, commensurability, etc.); it constitutes a shared platform to explore several aspects of value-driven behaviour (argumentation, negotiation, value adoption, etc.), and it provides experimental support for insights. Moreover, it also provides running examples of value-imbued systems as a side effect.

As will become clear with the next sections, expected outcomes follow three paths that we claim will be fruitful and significant: (i) formal results: on a cognitive theory of values, on institutional design, on value alignment assessment,

**Table 1.** Abstraction process for building simulators of public policies

| | Reality | Formal abstraction | Simulation |
|---|---|---|---|
| World | $W$ | $D$ | $W' \leftarrow \Sigma(I(p))$ |
| Policy | $\mathcal{P}$ | $P = \langle D, V, \Pi, S \rangle$ | $p \leftarrow i(P)$ |

on value-imbued agreement technologies; (ii) methodological guidelines for value representation, value-driven decision-making architectures, value-enhancing governance, uses of the simulations, etc.; and (iii) collateral uses like: methodology and workbenches for value-driven policy-models, frameworks for development and deployment of policy-support systems, experimental platforms for the study of social psychology and collective social phenomena.

## 3 A Conceptual Framework for ABM of Value-Driven Policy-Making

In loose terms, a public policy is a proposal to improve reality. In essence, *reality* ($W$) refers to a "fragment of the real world that is relevant" for the policy. This fragment is abstracted into a "model" ($D$) whose "implementation" ($I$) and running ($\Sigma$) produces a "simulated world" ($W'$) that corresponds with the relevant fragment of reality (Table 1).

A policy *proposal* consists of two types of elements: the "ends" that will be used to assess the *improvement*, and the "means" that will be used to achieve those ends. In other words, we define a (model of) public policy through four main elements: (i) a policy domain ($D$), which is that part of reality that the policy is intended to improve; (ii) the set of values ($V$) with respect one intends to assess the improvement; (iii) a policy-schema ($\Pi$), which is the core of the *proposal* (what to improve and how); and (iv) the population of stakeholders ($S$) that are involved in the policy *proposal*.

We are interested in building agent-based simulators of such policies. We shall call them Agent-Based *Value-Driven Policy*-making *Simulators* (VDPS, for short) (see [29]). These simulators are nothing more than the instantiation of a model of a public policy ($p \leftarrow i(P)$), its implementation ($I(p)$), and the computational environment where we run the simulations ($\Sigma(I(p))$) (Table 1). Theoretically, these systems could inform and support actual policy-making in the real world ($\mathcal{P}$).

All this may be expressed in more precise terms through the following working definitions:

**Working Definition 1.** *A **policy** is a tuple $P = \langle D, V, \Pi, S \rangle$, where*

- *$D$ is a policy-domain;*
- *$V$ is a set of values;*
- *$\Pi$ is a policy-schema;*

– *A is a set of stakeholders.*

**Working Definition 2.** *A **value-driven policy simulator** is a tuple* $VDPS = \langle M_{vdps}, I, \Sigma \rangle$*, where*

– $M_{vdps}$ *is the metamodel for such systems;*
– $I$ *is the implementation platform (e.g. NetLogo, Repast, etc.);*
– $\Sigma$ *is the simulation environment.*

The idea of a metamodel (see Appendix) is to provide *affordances* that enable the modeller to express those features and functionalities which facilitate the construction of the model.[1]

In our case, we want an abstract description of any policy, being this detailed enough to capture everything that is common to policies (and distinguishes them from other socio-cognitive technical systems (see Appendix)). The purpose is to instantiate a description, which then will be implemented in a platform and used for simulation. For that reason we explain the four elements we said constitute a policy: the *policy domain*, a *policy-schema*, the *values* in the policy, and the *stakeholders* involved.

First, the *policy domain* needs to be concrete enough (whatever is involved in, say, policies for the urban water management). Thus, we need an ontology (i.e., entities that constitute the world: water, households, tariffs, utility companies, etc.), and the semantics and pragmatics that describe their functioning (e.g., household water supply is measured in cubic meters, and the tariff is proportional to their actual water use). In other words:

**Working Definition 3.** *A **policy domain** D includes:*

1. *A **domain model**, given by*
   1.1. *The **ontology**, **semantics** and **pragmatics**.*
   1.2. *The **state of the world**, which is a finite set of variables that stand for crude facts.*
2. *An **institutional framework**, that contains*
   2.1. ***Policy instruments**: actions, norms, and activities (i.e., organised collections of actions performed by agents with specific roles)*
   2.2. *"**Shells**" for policy instruments (means to specify new actions, norms and activities).*
   2.3. ***Governance framework** to regulate those entities.*
3. ***Stakeholder roles**, defined in terms of capabilities, norms that govern their role-specific behaviour, and relationships among roles.*

Next, we need to model *values*. For that purpose, we start assuming we have a set of values, a set of agents, and the *policy-domain*:

[1] The model, its implementation and the resulting simulated world constitute a socio-cognitive technical system (see Appendix). The metamodel of Working Definition 7 characterises the sub-class of value drive policy-making simulators.

**Working Definition 4.** *Given a set of* ***domain values*** $V$*, a set of* ***domain actions*** $A$*, a set of* ***domain norms*** $N$*, a set of* ***domain activities*** $C$ *and a set of* ***domain roles*** $R$*. We distinguish between*

1. ***Social values****, which are those values that are involved in the definition of the policy (i.e., chosen by the policy-makers) and in the evaluation of the social outcomes. We assume that actions, norms, and activities are associated with these values:* $V_{social} = \{\langle a_i, v_a\rangle, \langle n_j, v_n\rangle, \langle c_k, v_c\rangle\}$*, where*
   - $v_a \subset V$ *are all the values involved in the* ***action*** $a_i$ *(inputs, outputs).*
   - $v_n \subset V$ *are all the values involved in the* ***norm*** $n_j$ *(preconditions, actions and postconditions).*
   - $v_c \subset V$ *are all the values involved in the* ***activity*** $c_k$ *(values in all actions and norms that define the protocols).*
2. ***Stakeholders' values****, which are those values that are held by individual agents and may be specific to some roles:* $v_r \subset V$
3. ***Value assessment conventions***
   - 3.1. *Define how to assess the degree of satisfaction of a set of values in a given state of the world.*
   - 3.2. *Define how to compare two states of the world (i.e., order states with respect to a set of values).*
   - 3.3. *Define how to assess the contribution of an action towards a set of values.*

A *policy-schema* is the way a policy is meant to be implemented: its *means* and *ends*. In more precise terms:

**Working Definition 5.** *A* ***policy-schema*** *is a tuple* $\Pi = \langle means, ends\rangle$*, where*

1. ***Means*** *are the ways to improve the state of the world, and include a set of* ***instruments*** *to influence and foster behaviour and make those means effective;*
2. ***Ends*** *are the description of the expected improvements, and include the set of* ***indicators*** *to measure the performance of the policy.*

For example, an *end* for urban water policy may be to guarantee a safe supply of water, which may be represented by a combination of *indicators* like water use (e.g. litres per person and day), water quality parameters (e.g. nitrates in milligrams per litre), etc. The *means* of the policy are interventions like controlling water supply and demand, which are implemented through corresponding *instruments* as higher tariffs for large consumers, smart water-meters that block supply when water use is abnormally high, etc.

The stakeholders are those agents—who may be individuals, groups of individuals, or organisations—who are capable of performing some actions in the policy domain and are subject to the means and ends instituted by the policy. Thus, a stakeholder is characterised by the capabilities that allow it to act in the policy domain (e.g., use water, influence and be influenced by neighbours, advocate for a change of tariffs, etc.) and by its decision model. We claim that this model involves values and other cognitive constructs like beliefs, motivations, personality, abstract reasoning, etc., which we lump into a "mindframe".

**Working Definition 6.** *A population of* ***stakeholders*** *is constituted by agents who are provided with:*

1. ***Capabilities****, i.e., actions that are able to be performed.*
2. ***Values*** *and* ***value assessment conventions****.*
3. ***Mindframes****, i.e., the decision model and inputs used (e.g., beliefs, resources, personality, etc.).*

When we put all these components together we can define a metamodel for VDPS as:

**Working Definition 7.** *A* ***metamodel*** *for value-driven policy simulators $M_{vdps}$ provides affordances to specify:*

1. *A* ***policy-domain****;*
2. ***Policy-schemas****;*
3. *A set of* ***domain values****;*
4. *A population of* ***stakeholders****.*

While the metamodel $M_{vdps}$ should aid modelling of any value-driven simulator, one may narrow the scope to a particular policy domain and be more explicit about the requirements of some metamodel components and define, with the more restrictive metamodel a "landmark class" of value-driven simulators that is still rich enough for theory and technological development. For example, if the class is restricted to urban water use policies, one does not have to deal with "all" possible values, but only to the ones that apply to the urban water domain; and therefore one does not need to model any possible action or event that affects the physical-socio-economic but only those means and ends linked with urban water use that have an impact in decision-making and assessments of the state of the world that are value-based.

While the aforementioned limitations do not entail any qualification of value theory, we propose to adopt two non-essential assumptions: consequentalism and commensurability of values. We claim that these assumptions still leave a rich enough class and also facilitate a systematic exploration of policy-simulations with other value-theory assumptions with the archetype defined through.

**Working Definition 8.** *A* ***landmark*** *class for value-driven policy simulators in a particular policy domain $D^*$ is the class of models that may be specified with the refinement of metamodel $M_{vdps}$ as a metamodel $M_{vdps}^{D^*}$ with the following components:*

1. *The domain submodel is restricted to the physical, economic and social environment of a particular $D^*$.*
2. *The list of relevant values involved in simulation are specific to $D^*$ and thus their operationalisation has to be $D^*$-specific.*
3. *Agent models need only be concerned with reasoning about $D^*$-relevant values.*
4. *Roles, actions, norms and activities are also $D^*$-specific.*
5. *The state of the world is represented by a* ***finite set of indicators*** *(domain variables and constants).*

6. *All values may be defined as a combination of indicators.*
7. *Value assessment for decision-making and for moral judgement is based in the* ***indicator-based definition of values*** *and value aggregation models that use those indicators.*

Note that both consequentalism and commensurability are rather natural for agent-based simulation. In the generic metamodel for VDPS $M_{vdps}$ we do not commit to a specific operationalisation of these assumptions but only that any model may involve that type of operationalisation. We propose to narrow further the scope of the archetype through a refinement of $M_{vdps}^{D^*}$ where the domain variables of each value (i.e. contextualised values are translated into indicators) and the value aggregation models are made explicit in the specification of a simulator.

**Working Definition 9.** *A* ***challenge*** *is a subclass of a landmark class through the refinement of* $M_{vdps}^{D^*}$*, where*

1. *For each value in* $M_{vdps}^{D^*}$*, the set of state indicators that are involved in its definition are chosen from the larger set of indicators of the contextualised values of the specific domain.*
2. *A set of value aggregation models are defined (for instance, an utility function).*

The idea is to work, for example, in an ideal city with an explicit definition of a state of the world, a ground ontology, social model and institutional framework (that may be expanded through policy instruments), a list of values that need to be operationalised (i.e. contextualised and translated into indicators), but this operationalisation is made explicit so it can be compared. Modest applications in the water domain have been used to study innovation-focused policies in farmer communities [30] (namely, innovation policy value is translated into adoption rate, promoted by means of subsidies, for farmers who are profit-driven), and modelling stakeholders' advocacy for policy shifts in the urban water domain [31] (stakeholders have different values and perceive them differently, and may propose policy instruments according to their evaluation of the state of the world).

## 4 Some Lines of Research and Development

One may organise a research programme around the policy-making archetype. One may start with the definition of a "challenge" test case whose domain submodel is properly implemented and publicly available. It may then be used to explore in a systematic way several questions that underlie the general value alignment problem.

The guiding light of this exploration is the dual aim of imbuing values in the system and imbuing values in autonomous agents. With that in mind one may split research and development along interleaving paths. The following list is but a rough one.

**What Is a Value in Policy-Making.** Find a useful definition of value. What are the differences between the use of the notion of value in policy-making and other contexts. What classes of values are involved in policy-making. Are value classes different for different stakeholders?

**What Values to Use?** What value sets have been proposed? What is the theoretical foundation? What is the backing? What methodology? What is the relationship between values and disvalues? What is the relationship among values? How are values ranked? Scope of applicability? Values for specific domains? Values for specific populations? How are values contextualised? How context alters the ranking? What contexts alter the rankings more?

**Value Operationalisation.** What is the meaning of a value? Methodologies? How is the value assessed? What are the conditions for assessment? What needs to be observable in the world to assess a given value? What actions affect or are affected by those observable entities? How are alternative definitions justified? Empirical backing?

**Values and Policy Ends.** What are the principal values in a policy? How are they captured in the ends of the policy? Indicators vs indexes? Value aggregation for ends? Can ends evolve? How ends relate with different values and disvalues? End commensurability?

**Values and Policy Means.** How are values linked to actions? How are values linked to norms? How are values linked to sanctions? How can one measure the significance of values, norms, rhetorical messages with respect to a value? What types of instruments are available for imbuing values? Are there some that are specific to classes of values? How to observe value effects with successions of actions? What are the costs of complying with a value? How to evaluate the quality of a means? How are alternative definitions justified? Empirical backing?

**Value Aggregation.** Alternative models? Advantages? Preferred contexts for application? Reliability? Testability? Robustness?

**Reasoning with Values?** What are the characteristics of value as a mental construct? What are the connections between values and goals, motivation, personality, beliefs, social environment, mind frames? Argue with values in mind?

**Policy Assessment.** When is a policy "good"? When is it effective? What features need to be assessed in a public policy? What means and ends are better for a policy? Existing policy assessment guidelines and practices? What values are meant to be imbued?

**The Archetype in the Wider Context of Policy-Making.** How are simulations part of the policy-making cycle? Where can simulation be used, how, and to what advantage? How policy simulation may be used as a support tool for policy making? How can a policy-making simulator be extended into a policy-support system?

**Ethical Aspects Involved in the Use of Value-Driven Agent-Based Simulators**. What to put into the model? How "good" is the model? What are

the responsibilities of the designer? What are the responsibilities of the user of the simulator? What is different in agent-based simulation and other type of models?

## 5 Closing Remarks

*Problem Definition.* We propose to frame the problem as follows: (i) study the process of policy-making as an instance of two aspects of value-alignment: instituting value-based governance, and programming value-driven behaviour; (ii) build an agent based simulator as an experimental environment; (iii) use socio-cognitive artificial agents as experimental subjects; (iv) make the problem operational by designing an archetype version of the simulator; (v) design a test-bed platform for challenges.

*Practical Matters.* One may proceed as follows: (i) use environmental policies as the archetype policy domain and water use policies for a landmark class; (ii) build the test-bed platform on top of existing domain models and agent-based modelling tools.; (iii) specify a challenge instance; (iv) provide a test bed platform with a precise (challenge) instance of the archetype where policy designers may test policy means and ends as well as agent architectures. Such instance would contain a working policy domain simulator, a data set to feed simulations and a core set of tools and shells; (v) define "game rules" that establish standard experimental scenarios.

*Significance.* As suggested in the Online Manifesto [16], being human in an hyper-connected era entails interacting in an augmented nature with autonomous intelligent entities that are artificial. We are witnessing the power of AI and fear we will loose control before we fully understand it [11]. That is why research and development that guarantees that artificial entities behave ethically is transcendental and urgent. We are convinced that a wise and timely way to approach this colossal task is to design a problem archetype that captures some key features of the value alignment problem, and simplify the archetype into challenges that facilitate and motivate the systematic exploration of fundamental questions.

**Acknowledgements.** This research has been supported by the CIMBVAL project (Spanish government, project # TIN2017-89758-R). The first author is supported with the industrial doctoral 2016DI043 grant of the Catalan Secretariat for Universities and Research (AGAUR), sponsored by FCC AQUALIA, IIIA-CSIC, and UAB.

## Appendix: Background

### Values

Values are constructs that are grounded on universal human needs [40]. Presumably, they are cognitive socio-cognitive constructions that articulate these

human needs as principles or standards of preference. Indeed, values are involved in motivation and goal-setting [28], and it has been suggested that they are also involved in political cognition [8,41], as they serve as moral intuitions for individuals [20]. For more practical terms, values play a role in the behaviour of agents [25,28].

Schwartz and Bilsky [40] provided an exhaustive definition of values: values are (a) concepts or beliefs, (b) about desirable end states or behaviours, (c) that transcend specific situations, (d) guide selection or evaluation of behaviour and events, and (e) are ordered by relative importance. They derived such cognitive notion of values from universal human needs (i.e. individual biological organism, social agent and interaction, and welfare of the community).

There is no consensus on the categories of values. Rokeach [35] developed the Rokeach Value Survey—on which Schwartz and Bilsky draw their primary work—to determine the value priorities of individuals and distinguished between *instrumental values* (i.e., related to modes of behaviour) and *terminal values* (i.e., desirable end-states of existence); and also between *individual values* (i.e., related to satisfying individual needs and self-esteem) and *societal values* (i.e, related to societal demands, since *supra-individual entities* (e.g., society, organisations, etc.) "socialise the individual for the common good to internalise shared conceptions of the desirable").

One of the most notable works on values, the Schwartz Theory of Basic Values, defines 10 sets of values that pursue a particular objective or goal [39]: Power; Achievement; Hedonism; Stimulation; Self-direction; Universalism; Benevolence; Conformity; Tradition; and Security. This theory has been used to study in political domains (e.g., voting behaviour [41]), and even has been subject of study to enhance their usability in public administration and policy studies (see, for instance, [45]).

**Cognitive Function of Values.** Values are largely stable social and internal cognitive constructs that represent individuals' moral intuitions and which guide social evaluation and action [20]. Accordingly, values play a role in perceiving the relevant fragment of the world, in evaluating the state of such, and in motivating responding action. It has been suggested that values are essential for the socio-political cognition of individuals (regarding social outcomes and public affairs) [8].

Generally speaking, decision-making within a particular context pose ethical dilemmas that present trade-offs between multiple values, revealing desirable but opposing outcomes. Noteworthy, any decision is value-laden because it reflects the hierarchy of values of the decision-maker.

When multiple values are involved in decision-situations, we say that values are made *commensurable* by means of *value aggregation models*. These decision-making components afford individuals to consider multiple values and solve value trade-offs, eventually making a decision. With this in mind, there are at least two relevant components in *value aggregation models*: (i) the value system, that defines the type of values considered (e.g., Schwartz, Rokeach, etc.); and (ii) the aggregation model (e.g., satisficing combinations, aggregation functions, etc.).

Usually, such models are implemented as aggregation functions that reflect a multi-criteria decision analysis (MCDA) [33]. Sophisticated mathematical protocols have been developed to generate *value functions* (see [1]) and *value hierarchies* (see [37]).

**Working Assumptions on Values.** We adopt the following assumptions about values to ground a working framework:

- **Cognitive understanding of values.** Values are constructs that serve as cognitive heuristics and moral intuitions of individuals, and therefore they guide perception, evaluation and decision-making in any context [20,40].
- **Commensurability of values.** Although one can say that values are incommensurable and cannot be measured on a common scale [33], we stick to the fact that individuals *act* and *make decisions*, which requires to solve ethical dilemmas and value trade-offs (e.g. for which we presume *bounded rationality* is crucial). Thus, values are, at least, *cognitively commensurable*.
- **A consequentialist view of values.** The focus of value-driven decisions is placed on their consequences, rather they nature and definition. In other words, the discussion is not about what a particular value is, but rather, given a definition of that particular value, whether actions promote or not that value.

**Values in Norms and in Actions** Values are related to norms. Values serve as guiding and evaluative principles that capture what is right and wrong, while norms are rules that prescribe behaviours and particular courses of action. Accordingly, norms are an "implementation" of what values express (either as a personal norm in the cognition of the individual or as a institutional norm in the social space).

According to our working framework, an action $A$ may promote a value $\alpha$ and demote a value $\beta$ depending on their consequences and how value $\alpha$ and value $\beta$ are understood. Alternatively, an action $A$ is aligned with a value $\alpha$ if the outcome improves the state of the world with respect to how that value $\alpha$ is understood. Following this approach, a norm $N$ is aligned to a value $\alpha$ when it prescribes actions that promote value $\alpha$ and prohibits actions that demote value $\alpha$.

**Policy-Making**

Public policies are plans of action that address what has been defined as a collective problem [13] in order to produce a desirable society-level outcome. Values play a role in policy decisions, as they are involved when defining public issues, desirable states of the world, and courses of action worth to be considered [8,43].

Ideally, policy-making cycle is often described as a linear cycle that includes agenda-setting, design, implementation, application, evaluation and revision (i.e., maintain, redesign, or terminate the policy). The truth is that policy-making is far more complex and uncertain than a linear process [8,9,38]. Noteworthy, policy decisions are usually made without enough information—which is not only based on scientific evidence, but also on habits and intuitions [9]—in a

space where multiple stakeholders are involved—who have competing values and interests, and mobilise diverse resources [13]—but they still may have substantial impact—whose consequences are often not totally foreseen [42].

**Policy Domains.** A policy domain is an abstraction of the reality that serve to draw the boundaries of the relevant fragment of the world to be considered when addressing public issues. In simple terms, it consists of going from a messy problematic situation to a structured well-defined problem, which affords to conceive policies to tackle it [21].

Paradigms are taken-for-granted descriptions and theoretical analyses that constrain the range of alternative policy options [23]. Paraphrasing Campbell [10], paradigms act as "cognitive background assumptions that constrain action by limiting the range of alternatives that policy-making elites are likely to perceive as useful and worth considering" when addressing public issues. These paradigms are supported by language and discourse, contributing to form "mental structures that shape the way we see the world" [24].

**Policy Ends and Indicators, and Policy Means and Instruments.** In simple terms, public policies are a set of values (i.e., what is valued by the society at large), ends (i.e., what state of the world reflects them), and means (i.e., how that state is going to be achieved).

Following this view, ends must be described clearly as objectives that are meant to be achieved by the intervention. The assessment of the degree of success may rely on indexes and indicators—either quantitative or qualitative—that stand for those end states and are computed from variables of the relevant world.

In the same vein, means aim to produce a change on the relevant world (typically, a behavioural change on target groups) so as to drive the system towards a desirable state of the world. They may be implemented with diverse instruments (e.g., financial, economic, regulatory, etc.).

**Policy Assessment Practices.** It is common to assess policies prior to their enactment (*ex ante* assessment). For instance, the European Commission refers to this process as Impact Assessment (IA), and considers it necessary when the expected economic, environmental or social impacts of interventions are likely to be *significant* (see [15]). The main steps of the process consists of analysing (i) the definition of the problem and boundaries and scales of the system; (ii) the policy ends and how are they going to be measured; (iii) the policy means and how are they going to be implemented; and finally (iv) the policy evaluation on which base the enactment, redesign, or termination decisions.

We distinguish between *effective* policies and *good* policies [32]. The former are those policies whose social outcome is consistent with the policy declared objectives. In contrast, the latter are those policies whose social outcome is "good" according to the values held by stakeholders.

### Agent-Based Simulation (ABS) for Policy-Making

Simulation is the imitation of a real-world process or system over time, and can contribute to policy assessment without disturbing the real social system

and committing resources [7], as well as identifying counter-intuitive situations. ABS use a type of computational models that are able to explicitly simulate the actions and social interactions of groups of individuals within an artificial environment, thus generating "artificial societies" [18].

With this in mind, agent-based simulation (ABS) has been acknowledged as a useful tool to support policy-making and *ex ante* policy assessment [19]. ABS contributes to reliably anticipate data that is not currently known [14], and can be combined with other ICTs to enhance their potential (e.g., data analysis and statistics, output visualisation, etc.). Although ABS is promising, several concerns have been posed, as it can backfire if used without proper precaution [5].

**Socio-cognitive Technical Systems**

Socio-cognitive technical systems (SCTS) are social coordination systems [2] that articulate on-line interactions of autonomous agents that are socially rational [26]. They are composed of two first class entities: a social space where all interactions take place, and agents that interact within that environment. One presumes that the social space has a fixed ontology (the *domain ontology*), that at any given time it is in a *state*—that is an instance of the Cartesian product of a finite number of domains, whose union is a subset of the domain ontology. The state of the system changes only as a result of an action that complies with the system regulations the moment it is attempted, or because an event that is compatible with those regulations takes place.

SCTS can be decomposed in three "views": $\mathcal{W}$ that is the fragment of the world that is relevant for the system, $\mathcal{I}$ an institutional representation of the conventions that define the system, and $\mathcal{T}$ the implementation of $\mathcal{I}$ that creates the on-line version of $\mathcal{W}$. The views are interrelated in such a way that an attempted action modifies the state of the system if and only if that action is admitted by the system interface, which in turn should happen if and only if the attempted action complies with the conventions established in $\mathcal{I}$ (and those conventions are properly implemented in $\mathcal{T}$). An admitted action changes the state of the world according to the conventions in $\mathcal{I}$ that specify the way the input is processed in $\mathcal{T}$. In the case of value-driven policy simulators, these three views correspond to the simulated world, the (abstract) model of the world and the implementation of the model.

In practice, the institutional specification ($\mathcal{I}$) is achieved by instantiating a *metamodel* that includes *ad-hoc* languages and data structures to represent key distinctive features (*affordances*) of a family of SCTS (e.g., crowd-based systems, electronic institutions [2], normative multiagent systems [4], second-order simulation [27]).

## References

1. Alarcon, B., Aguado, A., Manga, R., Josa, A.: A value function for assessing sustainability: application to industrial buildings. Sustainability **3**(1), 35–50 (2011)

2. Aldewereld, H., Boissier, O., Dignum, V., Noriega, P., Padget, J.: Social Coordination Frameworks for Social Technical Systems. Law, Governance and Technology Series, vol. 30. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-33570-4
3. Allen, C., Varner, G., Zinser, J.: Prolegomena to any future artificial moral agent. J. Exp. Theor. Artif. Intell. **12**, 251–261 (2000)
4. Andrighetto, G., Governatori, G., Noriega, P., van der Torre, L.W.N. (eds.): Normative Multi-Agent Systems, vol. 4. Dagstuhl Publishing, Saarbrücken (2013)
5. Aodha, L., Edmonds, B.: Some pitfalls to beware when applying models to issues of policy relevance. In: Edmonds, B., Meyer, R. (eds.) Simulating Social Complexity. UCS, pp. 801–822. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66948-9_29
6. Awad, E., et al.: The moral machine experiment. Nature **563**, 59 (2018)
7. Banks, J.: Handbook of Simulation. Wiley, Hoboken (1998)
8. Botterill, L.C., Fenna, A.: Interrogating Public Policy Theory. Edward Elgar Publishing, Cheltenham (2019)
9. Cairney, P.: The Politics of Evidence-Based Policy Making. Palgrave Macmillan, Basingstoke (2016)
10. Campbell, J.L.: Institutional analysis and the role of ideas in political economy. Theory Soc. **27**(3), 377–409 (1998)
11. Collingridge, D.: The Social Control of Technology. Palgrave Macmillan, Basingstoke (1981)
12. Asilomar Conference: Asilomar AI principles (2017). https://futureoflife.org/ai-principles/. Accessed 13 2019
13. Dente, B.: Understanding Policy Decisions. SpringerBriefs in Applied Sciences and Technology. Springer, Cham (2013)
14. Edmonds, B.: Different modelling purposes. In: Edmonds, B., Meyer, R. (eds.) Simulating Social Complexity. UCS, pp. 39–58. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66948-9_4
15. European Commission: Better Regulation Toolbox. https://ec.europa.eu/info/better-regulation-toolbox_en. Accessed 20 Mar 2019
16. Floridi, L. (ed.): The Onlife Manifesto, pp. 7–13. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-04093-6_2
17. Floridi, L., et al.: AI4People - an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach. **28**(4), 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
18. Gilbert, G.N., Conte, R.: Artificial Societies: The Computer Simulation of Social Life. UCL Press, London (1995)
19. Gilbert, N., Ahrweiler, P., Barbrook-Johnson, P., Narasimhan, K.P., Wilkinson, H.: Computational modelling of public policy: reflections on practice. J. Artif. Soc. Soc. Simul. **21**(1), 14 (2018)
20. Hitlin, S., Pinkston, K.: Values, attitudes, and ideologies: explicit and implicit constructs shaping perception and action. In: DeLamater, J., Ward, A. (eds.) Handbook of Social Psychology. Handbooks of Sociology and Social Research, pp. 319–339. Springer, Netherlands (2013). https://doi.org/10.1007/978-94-007-6772-0_11
21. Hoppe, R.: Heuristics for practitioners of policy design: rules-of-thumb for structuring unstructured problems. Public Policy Adm. **33**(4), 384–408 (2018)
22. IEEE: Ethically aligned design, version 2 (2017). https://ethicsinaction.ieee.org/. Accessed 13 2019
23. Jasanoff, S., Wynne, B.: Science and decision making. In: Rayner, S., Malone, E.L. (eds.) Human Choice and Climate Change, pp. 1–87. Battelle Press, Columbus (1998)

24. Lakoff, G.: Don't Think of an Elephant! Chelsea Green Publishing, Hartford (2004)
25. Miceli, M., Castelfranchi, C.: A cognitive approach to values. J. Theory Soc. Behav. **19**(2), 169–193 (1989)
26. Noriega, P., Padget, J., Verhagen, H., d'Inverno, M.: Towards a framework for socio-cognitive technical systems. In: Ghose, A., Oren, N., Telang, P., Thangarajah, J. (eds.) COIN 2014. LNCS (LNAI), vol. 9372, pp. 164–181. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25420-3_11
27. Noriega, P., Sabater-Mir, J., Verhagen, H., Padget, J., d'Inverno, M.: Identifying affordances for modelling second-order emergent phenomena with the $\mathcal{WIT}$ framework. In: Sukthankar, G., Rodriguez-Aguilar, J.A. (eds.) AAMAS 2017. LNCS (LNAI), vol. 10643, pp. 208–227. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71679-4_14
28. Parks, L., Guay, R.P.: Personality, values, and motivation. Pers. Individ. Differ. **47**(7), 675–684 (2009)
29. Perello-Moragues, A., Noriega, P.: Using agent-based simulation to understand the role of values in policy-making. In: Advances in Social Simulation – Looking in the Mirror (in Press)
30. Perello-Moragues, A., Noriega, P., Poch, M.: Modelling contingent technology adoption in farming irrigation communities. J. Artif. Soc. Soc. Simul. (in Press)
31. Perello-Moragues, A., Noriega, P., Popartan, A., Poch, M.: Modelling policy shift advocacy. In: Proceedings of the Multi-Agent-Based Simulation Workshop in AAMAS 2019 (in Press)
32. Perry, C.: ABCDE+F: a framework for thinking about water resources management. Water Int. **38**(1), 95–107 (2013)
33. Van de Poel, I.: Values in engineering design. In: Meijers, A.W.M. (ed.) Handbook of the Philosophy of Science, pp. 973–1006. Elsevier (2009)
34. Poel, I.: Translating values into design requirements. In: Michelfelder, D.P., McCarthy, N., Goldberg, D.E. (eds.) Philosophy and Engineering: Reflections on Practice, Principles and Process. PET, vol. 15, pp. 253–266. Springer, Dordrecht (2013). https://doi.org/10.1007/978-94-007-7762-0_20
35. Rokeach, M.: The Nature of Human Values. Free Press, New York (1973)
36. Russell, S.: Provably beneficial artificial intelligence. Exponential Life, BBVA-Open Mind, The Next Step (2017)
37. Saaty, T.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)
38. Sabatier, P.A.: Theories of the Policy Process. Westview Press, Boulder (1999)
39. Schwartz, S.H.: Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. In: Zanna, M.P. (ed.) Advances in Experimental Social Psychology, vol. 25, pp. 1–65. Academic Press (1992)
40. Schwartz, S.H., Bilsky, W.: Toward a universal psychological structure of human values. J. Pers. Soc. Psychol. **53**(3), 550–562 (1987)
41. Schwartz, S.H., Caprara, G.V., Vecchione, M.: Basic personal values, core political values, and voting: a longitudinal analysis. Polit. Psychol. **31**(3), 421–452 (2010)
42. Simon, H.A.: Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization. Macmillan, Oxford (1957)
43. Stewart, J.: Value conflict and policy change. In: Stewart, J. (ed.) Public Policy Values, pp. 33–46. Palgrave Macmillan, London (2009)
44. Susskind, J.: Future Politics: Living Together in a World Transformed by Tech. Oxford University Press, Oxford (2018)
45. Witesman, E., Walters, L.: Public service values: a new approach to the study of motivation in the public sphere. Public Adm. **92**(2), 375–405 (2014)