On Autonomy, Governance, and Values: An AGV Approach to Value Engineering*

Pablo Noriega and Enric Plaza

Artificial Intelligence Research Institute (IIIA-CSIC), 08193 Barcelona, Spain
pablo,enric@iiia.csic.es

Abstract. In this paper we show how to approach the engineering of values in social human-AI systems by looking into the interplay of three notions: autonomy, governance and value (AGV). We propose a particular characterisation of the Value Alignment Problem based on this approach. We use an example to illustrate how the values can be engineered in order to solve the VAP in this example. Based on these elements we advance some arguments in favour of framing a theory of values that may apply to the governance of social systems that involve artificial autonomous entities as well as humans.

Keywords: Engineering values · Value alignment · AI-inspired theory of values · Online Institutions · Value-driven policy design

1 Introduction

In the past few years, public opinion has raised concerns about AI. Most are based in the disruptive nature of AI, which arguably resides in the autonomy of AI artefacts. In order to abate these concerns and derive desirable outcomes of that autonomy, the AI community has taken up the challenge of engineering values in AI systems.

We propose to approach this challenge —in the spirit that drove the original efforts in AI, as we discuss below— by examining the interplay of the notions of autonomy, governance and value (AGV) in the context of social human-AI systems.

The point of this paper is to outline this AGV approach, argue for its relevance for value engineering in particular, and advocate for this approach as a research area in itself.

The paper is organised as follows: We start with a characterisation of the Value Alignment Problem (VAP) and the notion of value engineering is Sec. 2.

* 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.

Sec. 3 discusses the assumptions that underlie our AGV approach proposal and Sec. 4 illustrates how it can be applied to that particular case of agent-based simulation of a public policy. Sec. 5 suggests the type of research questions that can be explored with the AGV approach. We put these elements in a wider perspective by proposing an AI-inspired theory of values in Sec. 6.

2 An AGV characterisation of the VAP

The objective of AI has been defined as the design and construction of autonomous artefacts [15] and Russell himself has argued that the field of AI should take the "the design of systems that are provably aligned with human values" [14] as its main challenge (the so called *Value Alignment Problem* or VAP).

We postulate that **the VAP is a design problem.** The challenge is to design and build systems where values are an essential design consideration. We refer to the process of making the notion of value operational for value-aligned systems as *value engineering*. The following assumptions make the VAP design process more precise.

Vap.1 *Values are engineered into a specific system.* Consequently, (i) the system belongs to a certain domain (health, e-commerce, mobility); (ii) it involves specific stakeholders who are involved in the design of the system and (iii) abstract human values are contextualised to the domain of application of the system and the system stakeholders

Vap.2 *Values can be engineered in a three stage cycle*: values selection, embedding and assessment.

Vap.3 *Values are explicit.* Design is meant to be aligned with a specific set of values, thus each value needs to be interpreted, instrumented and assessed for the specific system, domain and stakeholders.

Vap.4 *The alignment of a value can be assessed in an objective way.* The expression "provably aligned" used in the definition of the VAP, need not be interpreted formally in proof-theoretic theoretic terms. We assume only that there is an objective way of determining to what degree a system is aligned with a value.

Vap.5 *Value aggregation.* Several values may be intended to apply. Thus, alignment is to be assessed with respect to the simultaneous application of the set of values.

3 Value engineering in social human-AI systems

In this section we make explicit some added assumptions that further constrain the characterisation of the VAP in order to make value engineering feasible. Notice that these assumptions reflect the interplay among the notions of Value, autonomy and governance in the context of social human-AI systems

Assumptions about values We adopt a rather standard interpretation of values (f.e., [16, 13]) that can be summarised in the following six assumptions:

Vl.1 Values motivate and legitimise goals.
Vl.2 Values determine preferences between states of the world.
Vl.3 Values are contextual.
Vl.4 Values may be in conflict.
Vl.5 Actions change the state of the world. Thus actions contribute to the achievement of values (promote, demote, protect).

Assumptions V.1 and V.3 allow us to focus into goals and specialise these to a particular system and its stakeholders. Assumption V.3 supports the objective assessment of the degree to which a value is being supported (Claim 3), and together with Assumption V.4 provide support for different ways of assessing value alignment. Finally, Assumption V.5 is used for the identification of instruments that support or promote values.

Assumptions about the social human-AI system Because we make the VAP a design problem and restrict value engineering to a particular system we narrow the scope of this engineering to social systems where autonomous entities engage in collective action within that system. This entails the following assumptions:

SS.1 There are two first class entities: the social system itself and the entities that interact within it. There is an inside and and outside of the social system. The social system controls what entities are active within it.
SS.2 There can be human as well as autonomous artificial entities active within the social system. We refer to these as "participating agents"
SS.3 Participating agents are autonomous in the sense that it is they who decide on their own whether to enter and leave the system and what actions they attempt while being active in the system.
SS.4 The system enables capabilities and establishes and enforces constraints to coordinate the actions of entities that interact within it.
SS.5 Only those attempted actions that comply with the system enabled capabilities and constraints can have an effect on the system.

Assumptions for engineering governance in social human-AI systems. We focus our attention on the value-driven governance of collective interaction, namely the governance of social systems that involve humans as well as artificial autonomous entities [1]. This translates the VAP into a two-fold governance problem as the following assumptions clarify:

Gov.1 Values can be engineered into the decision-making process of artificial autonomous agents who participate in the system.
Gov.2 Values are engineered into the coordination mechanisms that govern individual behaviour of autonomous human or artificial agents who interact within the system.

These two assumptions become operational during the *embedding* stage of the value engineering cycle, either into the decision-making architecture of artificial agents, or as governance mechanisms for collective interaction. How these two types of embedding can be engineered is illustrated in the example of the next section, however, there are some canonical ways to embed values in a large well-defined class social human AI system where some extra features hold (see [9]).

The Objective Stance. This is the interpretation of **Vap.4**, and is at the core of the three-stage engineering cycle. In order to make **Vap.3** operational, and based on *V.2*, we postulate the following two assumptions :

OS.1 The state of the world is observable
OS.2 The satisfaction of a value can be assessed through the state of the world.

Assumption *Os*.2 in combination with *V.1* and *V.2*, supports the key heuristic of associating values with goals. Once this association is made, the interpretation of the assessment of the satisfaction of goals all the way to defining different ways of assessing value alignment is straightforward. One may attempt to shy away from a strict notion of goal as a finite set of indicators and one may still hold to the Objective Stance; for instance, identifying a value with a particular state of the system that ought to be reached or avoided at any cost. In such case, value satisfaction is a binary decision and value instrumentation would be programmed into the system as a procedure that either always produces or always avoids the critical state.[1]

Stages on the value engineering cycle. The cycle of value engineering three stages: Choosing values, embedding values, and assessing value alignment.

1. **Choice (and contextualisation) of values.** The purpose of this stage is to focus on the values that are relevant for the intended purpose of the system and for its stakeholders. The outcome is a list of *value labels* that capture the intuitive understanding of the relevant values.
2. **Value embedding** Consists of making precise the meaning of each label, so that one can objectively assess to what degree it is being supported and to identify the means that will make the agents' behaviour and the collective space align with that value.
   (a) *Value interpretation* Turn each label into observable features that can be implemented in the system either as as a function of the state of the world (a goal) or as a critical state that ought to be reached or avoided. This interpretation conditions how the value may be instrumented and how value satisfaction can be assessed.

[1] The paraphrasing of *Vap*.3 as *OS*.1 and *OS*.2 is the least committed expression of *Vap*.3 we have been able to find. The main advantage is that, depending on the way one makes *1* and *2* operational, one may capture different interpretation of the notion of value, the means (instruments) to promote a value, and how one can assess value alignment [9].

(b) *Value instrumentation.* In practice, and because of Assumption *SS.*4 (the system enables and constraints agents actions) the means to guide the performance of a value aligned system towards achieving a value are designed in terms of those *actions* that affect the goals or critical states mentioned above. Because of assumptions *Gov.*1 and .2, the means to conduce the performance of a system towards achieving a value are slightly different.

i. *Autonomous agents.* Values are implemented into the decision-making architecture of an autonomous agent as: automatic behaviour, learned behaviour or value-based reasoning.

ii. *Governance of collective interactions.* Values are implemented as: (i) enabling or inhibition of potential actions, (ii) norms, conventions and artificial constraints that guide goal and critical state associated actions, (iii) information that becomes available to participating agents and may affect their decision-making, and (iv) deployment of participating agents whose behaviour is endorsed by the system (e.g. norm-enforcers).

3. **Alignment assessment.** This is a key design decision and, as mentioned above, can be made operational in several ways. There are some examples in the next section. It can be organised as follows:

(a) *Value-satisfaction function* that maps each state of the world to a degree of satisfaction of the goal (that stands for a value). Similarly, when a value is defined in terms of critical states (e.g., the degree of satisfaction for critical states is "acceptable/inacceptable", and maybe "indifferent" for all other states).

(b) *Value aggregation function* when several values are simultaneously involved *(Vap.5)*. This aggregation function should take into account conflicts among values *Vl.4*.

4 An application of the AGV approach: value driven policy design

Policy design as value-driven process In very loose therms, a policy is designed in order to improve the current state of affairs. 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 [4]. In this article we will apply this policy design to the Urban Water Use (UWU) domain (reported in [11]). Policy design assumes tehre is a policy domain (here urban water use) and policy stakeholders (city government, households, utility companies,etc).

Values determine, what are the actions that stakeholders prefer to take, and 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.

Policy design is a complex problem with several variables with complex interactions, involving several (often conflicting) motivations and interests and requiring factual as well as ethical decisions (cf. H. Simon [17]). As for other complex problems of this sort, simulation is a reasonable methodological approach to policy design [3, 10]. In fact, agent-based simulation (ABS) is particularly appropriate for policy design because it separates design concerns in the modelling of individuals (as autonomous agents) and in the modelling of collective action (with its coordination mechanisms).

ABS for policy interventions can be seen as a particular form of the VAP: It is a design process with the two main problems pf embedding values in a social human-AI, and assessing that the behaviour of the system is objectively aligned with those values. The following paragraphs discuss the key modelling components and the process of engineering values in the ABS. For a more detailed discussion see [6].

The social human-AI environment One needs two abstractions to represent the social system (SS.1) of the policy domain: a physical representation (Φ) of the natural constraints and capabilities of relevant part of the world (relevant entities, available actions, causal relations and observable effects) and an institutional abstraction (Ψ) of the artificial constraints that govern stakeholder activity, and their enforcement mechanisms. Individuals are modelled as autonomous agents (see below).

The state of the world In the simulation, the state of the world is a finite set of parameters or *indicators* in Φ. This state evolves only through actions that comply with Ψ and events that are recognised in Φ.

Modelling autonomy In our approach, *stakeholder modelling* amounts to some assumptions about the population of agents and the modelling of their decision-making. The core modelling assumes that an agent takes a policy-enabled action only when opportunity, capability and motivation concur.

On the other hand, *social governance* is modelled through the institutional constraints mentioned above. These constraints can be represented in several formalisms: from ad-hoc coordination protocols (like finite-state machines) to full normative multiagent systems (like those reported in [1, 2, 5])

Value Engineering: As we described in the previous section, value engineering can be divided in three stages, that we will now apply to the domain of urban water use.

1. **Value selection.** Identifying the values that are appropriate for the policy domain and the stakeholders whose values ought to be represented in the policy that is being designed. Urban water use should 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 embedding.** This stage addresses two issues:
   (a) *Making values observable.* That is, turn a label that stands for an abstract value into a feature that is measurable. Namely a goal (an indicator in the state of the world or a combination of indicators) 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 UWU, the goal to reduce individual water consumption to a certain average *per-capita* volume is an indicator for the consensual value of sustainability; a particular household may not even care about sustainability.
   (b) *Instrumenting values in a policy intervention.* Instrumenting involves choosing the 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, for example, to achieve sustainability, one may want to promote the adoption of water saving devices in households 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.
   Instrumenting values in a participating agent. In this case, values are instrumented in a combination of reactive behaviour, conditioned behaviour (learned) and value-driven reasoning [12]. When drawing a plan, the agent chooses actions that eventually lead it to states that are in alignment with its own values (see Fig. 1)

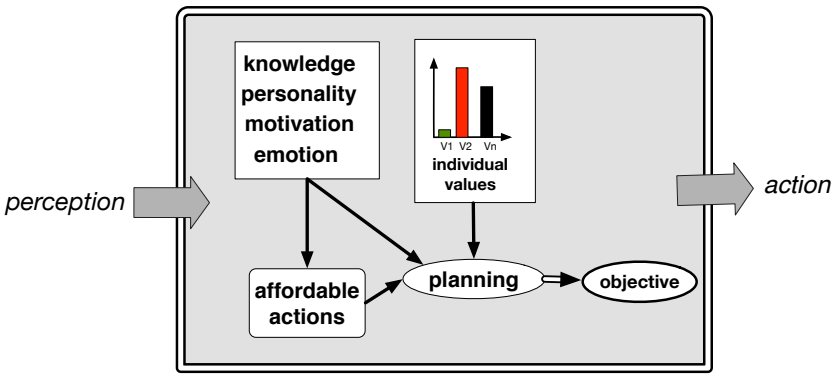

Fig. 1. Value-driven decision model for an individual household (from [6])

3. **Assessment of value alignment** We propose three value alignment functions to perform this assessment: effectiveness, adequacy and acceptability.
   (a) *Effectiveness* determines the degree to which a policy intervention is aligned with the policy values. In practice it measures the degree to which the policy intervention fulfils the policy objectives and is defined by the aggregation of value satisfaction functions for the policy goals. Analogously, one can also measure *effectiveness* with respect to the specific values of a given stakeholder.
   (b) *Adequacy* determines the trade-offs in direct and indirect costs of the instruments used in the policy intervention. This allows to compare equally effective interventions. Adequacy can be applied both to consensual or individual costs with respect to consensual effectiveness.
   (c) *Acceptability* (for a stakeholder) aggregates the combination of adequacy and effectiveness alignments for that stakeholder.

Policy interventions. A policy intervention is a collection of *policy instruments* whose effects on the state of the world are to be assessed (e.g., incentives to adopt water saving technologies, new water-treatment plants). Note that, as we mentioned above, such instruments are the way that values are embedded in the governance of the policy domain (i.e., the social human-AI system).

Using ABS to find a good policy intervention One can implement an agent based system that captures the features described above and then use agent-based simulation to design a policy intervention. One needs to develop the model to achieve its adequacy for policy design purposes (forecasting, epistemic and rhetorical properties) and then calibrate it so that simulation runs serve to elucidate policy-design decisions and eventually choose a policy intervention that would ideally be deployed. Without going into detail, this process is describe as a testing cycle that is summarised in Fig. 2

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 Research topics

The two previous sections outline how one can articulate an AGV approach to value engineering and how it can be put to work. In this section we identify some research topics that suggest how the AGV approach for value engineering may be developed.

These topics are roughly organised in three conventional categories. The next section puts these research topics in a wider perspective.

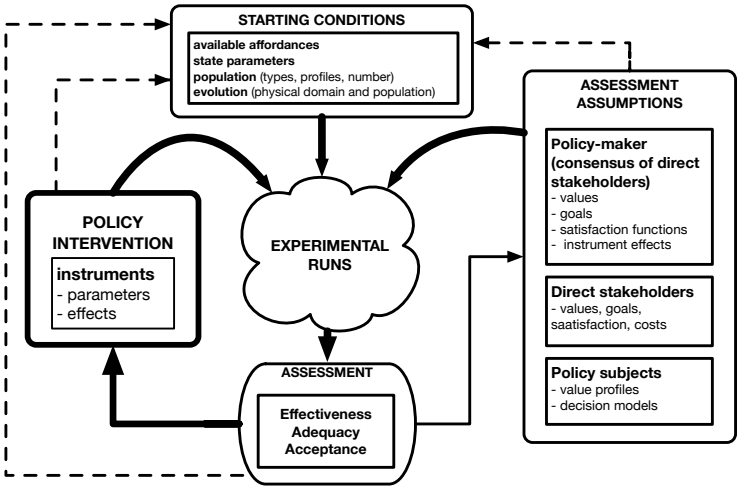

Fig. 2. The experimental cycle for testing and choosing policy interventions (from [6])

5.1 Meta-ethics

These are conceptual, theoretical or scientific aspects of the interplay of autonomy, governance and values in the context of social human-AI systems);

- What types of autonomy can be designed into artificial systems
- Values and governance. Relationships with compliance and legal principles
- What are values for? (what is new or different for design, for theory, for innovation...)
- What other notions are involved beyond AGV?
- Can there be artificial agency? the problem of intention, the level of abstraction, the problem of evil
- What cognitive features are linked with values (awareness, emotion, reasoning, planning, will, ...) and what is their role in value engineering?
- What are the differences between values for individual behaviour and for collective interaction? What is unifying and what should discriminate?

5.2 Normative ethics

Topics associated with the development of artifacts that can embed value driven governance in artificial autonomous systems and in social human-AI systems.

- How to choose values. What values are appropriate for a particula AIS
- How to validate an interpretation. How to link values and goals
- What is a value satisfaction function? What properties, uses, needs.... should a value assessemnt function provide? Are they all utility functions? How simpler can they get in order to achieve what?

- Forms of value aggregation. Commensurability? The problems with utilitarianism.
- How can other theories of value be instrumented?
- What forms of governance and compliance may be used to engineer values?
- The evolution of the instrumentation of values in a given system. Examples and methodologies for experimental ethics.

5.3 Applied ethics

These topics correspond to strategic, educational, sociological and regulatory issues that have to do value-driven governance of actual social human-AI systems.

- What about, "legal" aspects of value imbuing? standards, guidelines, regulations
- Standards for value-compliance
- Compliance as code versus value-driven governance of social human-AI systems.
- The discussion of ethics and AI. What is the role of value engineering in that debate? How would an AI inspired theory of values affect the debate?
- Heuristics for building vale-aligned systems
- Add-ons to guarantee that an app becomse value-aligned
- Assessement of value alignment in games, government apps, health, ...

6 AGV Beyond value engineering

We propose: to develop the AGV approach we have been so far discussing into a more general *AI-inspired Theory of Values*. We would like this theory to apply to the understanding of AI as the design and construction of artificial systems that have some form of *autonomy*. We are still motivated by the realisation that such autonomy needs to be *governed* and, finally, that the notion *value* may inspire interesting ways of harnessing potential AI developments.

Claim. The AI-inspired theory of values we propose includes value engineering and the VAP but has a wider scope.

We understand that the innovative aspect to be developed is that such a theory of values takes into account artificial autonomy, something that has not been addressed in other theories of values. It is new because it inspects ethical notions that are meant to apply to artificial autonomy (new types of entities) and because one intends to engineer moral aspects into artefacts in order to govern their behaviour.

Claim: The theory we propose is AI inspired in five specific aspects:

We propose:

1. **To study a leading abstract notion and the interplay of a small set of intimately associated notions:** intelligence versus value; and rationality, cognition, computer modelling, versus value-driven behaviour, autonomy, governance and collective action in social human-AI systems.
2. **A science of the artificial** that studies abstract problems and develops theoretical foundations with the goal of designing and engineering artefacts.
3. **Building on the tradition, technologies, practices and results of classical AI**: bridge the gap between theory and engineering; simulation and model building; methods like means-ends analysis and problem decomposition; lines and technologies like learning, reasoning, multiagent systems, etc.
4. **Select paradigmatic problems** to study salient features and develop approaches, methodological guidelines and constricts that may be used to understand aspects of the four-notions interplay (values, governance, autonomy and collective action). For example, policy design, value-driven design of online systems, autonomous vehicles, ...
5. **A multidisciplinary approach:** Discuss and explore insights and contribution of these AGV notions in different disciplines: take into account descriptive and operational notions of value, governance and autonomy for instance, motivational and cognitive aspects addressed in social psychology, the interplay of autonomy, rights and governance from political science and law, and notions of preference and value-based analysis from management science and behavioural economics and the ethical and aesthetic discussion of what is "good" and what are the "right" actions to take from a subjective moral perspective.

.

Claim: We expect such effort will report results along three wide types In analogy to three views of AI.

1. From the *mimetic* role of AI (to understand reality by building and studying models of reality): express in AI terms intuitions, insights and distinctions connected with the key concepts.
2. From the *prosthetic* role of AI: engineering AI enabled systems that reflect ethical concerns.
3. From the *symbiotic* environment where artificial autonomous entities engage with natural autonomous entities: explore specific contexts, elaborate guidelines, examples and concerns; propose specific contents for education, good practices and regulation.

Claim: We envision that this proposal will have impact along the following lines:

1. Bring coherence to a large corpus of related research in the AI community and elsewhere.
2. A crisper characterisation of interesting problems that, as suggested by our comments in Sec 5, can be organised in:

(a) meta-ethics: agency, role of values, characterisation of values,...;
(b) normative ethics: engineering values in AI enabled systems;
(c) practical ethics: regulation; education; good practices

7 Closing remarks

1. A particular example. The agent based modelling of value driven policy design (Sec.4) can be seen as a paradigmatic version of the VAP and the use of a typical AI-inspired experimental treatment for its exploration. The model includes the five core components *(Vap.*1-5); values play a key role in the definition of the system (*Vl.1-6*; it models a hybrid social system with artificial and human autonomous entities (*SS.*1-4); addresses the two levels of individual and coordination governance (*Gov.* 1-4), by construction of the simulation supports the Objective Stance and it follows the three-stage value engineering cycle (Vap.2) to in the end produce a properly value aligned system.[2]

2. A class of value aligned systems. The assumptions for AGV approach to value engineering discussed in Sec. 3 can be complemented with specific additional assumptions to characterise a class of online systems that are by construction value aligned [9]. This extension includes assumptions of two types. First, it limits the class of potentially value aligned systems to the class of *Online Institutions* [5]. Second, it assumes a design methodology, *Conscientious Design* [7], that structures and validates the value engineering process. The discussion of this extension in [9] and a preliminary version in [8] include concrete heuristics that provide alternative means of making the extension operational.

3. A principled approach to the VAP. The extension just discussed provides the starting components of a principled approach to the VAP along he topics mentioned in Sec. 5 and within the cope outlined in Sec.6. Such approach would include: Conceptual distinctions, constructs and properties to define the research space; as well as methodological guidelines and specific heuristics that guide the development of specific systems and apply the notions in (i) to AI-based systems in general. This principled approach would provide solid support for AI-inspired work in meta-ethics and for normative and educational proposals for the ethical use of AI.

Reviewer 1 (PC)

1. SUMMARY. Please briefly summarise the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here). *

The paper discusses the problem of value engineering in socio-technical systems.

2. DETAILED FEEDBACK. Please provide detailed, constructive, feedback to the authors concerning the strengths and weaknesses of the paper. Please take

[2] In fact nothing prevents participatory simulation where some stakeholders are in fact humans.

into consideration the novelty, soundness and impact of the work, the clarity of the presentation, the reproducibility of results, and ethical considerations (if any).

Minor comments *page 1 interplay of-¿ interplay among*
page 2 values determine, what -¿ values determine what
Applied ethics section (the heading appears twice)

3. OVERALL EVALUATION. Please provide your overall evaluation of the paper.

Accept 4. REASONS FOR THE EVALUATION. Please provide a summary of the reasons behind your overall evaluation. *

The discussion is relevant and interesting and the application to the policy design problem is also a nice inclusion in a paper of this nature.

Reviewer 2 (PC)

1. SUMMARY. Please briefly summarise the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here). *

The paper presents a framework to embed values in social human-AI systems that is based on implementing the concepts of autonomy, governance and value (AGV), with a focus on operationalising values to characterise the value alignment problem (VAP) quantitatively. From the statement of VAP principles, AGV assumptions are derived. Then the framework is illustrated with an implementation example (from [6]) and applications to policy making are references. Finally, some research questions stemming from the AGV approach are presented and the major claims are discussed in the context of theory of values for AI systems.

2. **DETAILED FEEDBACK.** Please provide detailed, constructive, feedback to the authors concerning the strengths and weaknesses of the paper. Please take into consideration the novelty, soundness and impact of the work, the clarity of the presentation, the reproducibility of results, and ethical considerations (if any). *

The assumptions from VAP to AGV are reasonable, concisely explained and well cohesive with one another. The example from section 4 nicely clarifies the approach.* In the initial stage of Value Engineering for the example in section, values are *explicitly designed in accordance with the stakeholders, making the implementation of values not exportable to other systems, at least in an obvious way.* It would be informative to **have a comment on whether different approaches have been attempted or if they are even feasible** (e.g., are there examples of value inference from existing systems?). **Requires proofreading to correct many typographic mistakes and inconsistencies checks.**

3. OVERALL EVALUATION. Please provide your overall evaluation of the paper. *

Accept 4. REASONS FOR THE EVALUATION. Please provide a summary of the reasons behind your overall evaluation. *

High-level principles, their ramification and illustration with examples provide useful guidance on how to approach value engineering in AI. **This research**

subject is becoming more relevant with the advent of powerful AI systems whose capabilities are not well understood and require better insight about their control, interpretability and the extent of their agency.

References

1. Aldewereld, H., Boissier, O., Dignum, V., Noriega, P., Padget, J. (eds.): Social Coordination Frameworks for Social Technical Systems, Law, Governance and Technology Series, vol. 30. Springer International Publishing (July 2016). https://doi.org/10.1007/978-3-319-33570-4, http://opus.bath.ac.uk/50167/, dOI: 10.1007/978-3-319-33570-4, ISBN: 978-3-319-33568-1 (hardcover), ISBN: 978-3-319-33570-4 (ebook)
2. Andrighetto, G., Governatori, G., Noriega, P., van der Torre, L.W.N. (eds.): Normative Multi-Agent Systems, vol. 4. Dagstuhl Publishing (2013)
3. Gilbert, N., Ahrweiler, P., Barbrook-Johnson, P., Narasimhan, K.P., Wilkinson, H.: Computational modelling of public policy: Reflections on practice. Journal of Artificial Societies and Social Simulation **21**(1), 14 (2018)
4. May, P.J.: Policy design and implementation. In: Peters, B., Pierre, J. (eds.) The SAGE Handbook of Public Administration, pp. 279–291. SAGE Publications, 2nd edn. (2012)
5. Noriega, P., Padget, J., Verhagen, H., d'Inverno, M.: Anchoring online institutions. In: Casanovas, P., Moreso, J.J. (eds.) Anchoring Institutions. Democracy and Regulations in a Global and Semi-automated World. Springer ((in press))
6. Noriega, P., Plaza, E.: The Use of Agent-based Simulation of Public Policy Design to Study the Value Alignment Problem. In: Casanovas, P. (ed.) Artificial Intelligence Governance, Ethics and Law (AIGEL). pp. 60–78. CEUR Workshop Proceedings, CEUR (In press)
7. Noriega, P., Verhagen, H., Padget, J., d'Inverno, M.: Ethical online AI systems through conscientious design. IEEE Internet Computing **25**(06), 58–64 (nov 2021). https://doi.org/10.1109/MIC.2021.3098324
8. Noriega, P., Verhagen, H., Padget, J., d'Inverno, M.: Design heuristics for ethical online institutions. In: Ajmeri, N., Morris Martin, A., Savarimuthu, B.T.R. (eds.) Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV. pp. 213–230. Springer International Publishing, Cham (2022)
9. Noriega, P., Verhagen, H., Padget, J., d'Inverno, M.: Addressing the value alignment problem through online institutions. In: Fornara, N. (ed.) Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVI. p. in press. Springer International Publishing, Cham (in press)
10. Perello-Moragues, A., Noriega, P.: Using agent-based simulation to understand the role of values in policy-making. In: Advances in Social Simulation. pp. 355–369. Springer (2020). https://doi.org/https://doi.org/10.1007/978-3-030-34127-5˙35
11. Perello-Moragues, A., Poch, M., Sauri, D., Popartan, L.A., Noriega, P.: Modelling domestic water use in metropolitan areas using socio-cognitive agents. Water **13**(8) (2021). https://doi.org/10.3390/w13081024, https://www.mdpi.com/2073-4441/13/8/1024
12. Rangel, A., Camerer, C., Montague, P.R.: A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience **9**(7), 545–556 (2008). https://doi.org/10.1038/nrn2357, https://doi.org/10.1038/nrn2357

13. Rohan, M.J.: A rose by any name? the values construct. Personality and Social Psychology Review **4**(3), 255–277 (2000)
14. Russell, S.: Of myths and moonshine. EDGE (2014), accessed: 2021-12-01
15. Russell, S.: Of Myths and Moonshine. A conversation with Jaron Lanier, 14-11-14. The Edge (November 2014), https://www.edge.org/conversation/the-myth-of-ai26015, [Online] Retrieved 12 december 2022
16. Schwartz, S.H.: Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In: Advances in experimental social psychology, vol. 25, pp. 1–65. Elsevier (1992)
17. Simon, H.A.: Fact and Value in Decision-making. In: Administrative Behavior: A study of decision-making processes in administrative organization. The Free Press, 4th edn. (1997)