Arguments for Framing the Value Alignment Problems with Online Institutions (Full)

Pablo Noriega[1] $^{[0000-0003-1317-2541]}$, Harko Verhagen[2] $^{[0000-0002-7937-2944]}$, Julian Padget[3] $^{[0000-0003-1314-2094]}$, and Mark d'Inverno[4] $^{[0000-0001-8826-5190]}$

[1] CSIC-IIIA, 08193 Bellaterra, Spain
[2] Stockholm University, 114 19 Stockholm, Sweden
[3] University of Bath, Bath, BA2 7AY, U.K.
[4] Goldsmiths, University of London, London, SE14 6NW, U.K.

Abstract. Online Institutions (OIs) are a specific kind of governed hybrid multiagent system, where norms guide human and software behaviour, and the norm compliance of each agent is observed. In this paper, we discuss how the characteristic features of OIs give rise to some key assumptions about the nature of values that can be used to engineer values into OIs. By outlining these specific features of values, we are one step closer to a general method for engineering value-driven governance into Artificial Intelligent Systems (AIS). That is, even though our focus is on a sub-class of AIS, there are ramifications for the design of general AIS where there is a goal to evidence how its operation, the affordances allowed, and behaviours supported by its constituent agents, embody a given set of values.

Keywords: Engineering Values, Value Alignment, Online Institutions, WIT Design Pattern, Conscientious Design

1 Motivation and Background

The objective of AI has been characterised as the design and construction of artificial autonomous entities. Arguably, such autonomy is the source of the most significant contributions of AI to society but also of its most significant concerns. One way to contend with artificial autonomy is to incorporate ethical concerns into the design and construction of artificial systems. In particular, to conceive ethics as a means of controlling that autonomy. Stuart Russell coined this challenge as the Value Alignment Problem (VAP) and stated: *to build systems that are provably aligned with human values* [11]. The Value Alignment Problem can be understood an engineering challenge that needs a rigorous approximation to the notion of "value" if one intends to evidence the degree to which an artificial system objectively aligns with a set of values. We refer to the notion of engineering such systems with a view to subsequently evidence their operation as "imbuing AI systems with values".

This paper is about the VAP challenge. It is one step in establishing the foundations for a principled approach to VAP that involves a precise delineation of the problem, to postulate the relevant conceptual distinctions and then develop notions and constructs that are particular to the problem, the properties of those constructs, and the heuristics and methodological guidelines that can support the design and deployment of value-aligned artificial autonomous systems. We propose here framing VAP for a particular class of artificial intelligent systems: online institutions (OI). We claim that the value alignment of OIs involves the same requirements as the full VAP but that, by definition, OIs have some particular features that allow a crisper characterisation of how values can be imbued in an OI, and provide an objective way of assessing the alignment of the OI to those values.

This paper is simply an argument for that claim organised in three parts. First, to set the terms of the argument, in Sec. 2 we present a broad motivation for online institutions and in Sec. 3 discuss their most relevant features in intuitive terms. Next, in Sec. 4 we make explicit some assumptions about values that can be predicated on electronic institutions. Finally, in Sec. 5, we enumerate specific heuristics that illustrate how these assumptions support the dual empirical problem of imbuing values in a system and assessing that the resulting system is objectively aligned. The final section gives some context for future work.

We choose this narrow focus because this paper is another step in our principled approach to the VAP, and those technical details and their contextualisation that complement the argument we present here can be filled by with what is discussed in four previous publications. (i) In "A Manifesto for Conscientious Design" [5] we outlined a research programme for value-driven design of artificial intelligent system; (ii) "Anchoring Online Institutions" [4] contains a more systematic presentation of the contents of sections 2 and 3. (iii) In "Design Heuristics for Online Institutions" [7] we presented our proposal for a principled approach to VAP and discuss in some detail the motivation, background and core elements of the proposal. (iv) Finally, in "Design Heuristics for Online Ethical Online Institutions" [6] we discussed the value operationalisation process and some heuristics for how to attack the process. [(i)]

2 An intuitive view of OIs

Online Institutions are inspired by a set of overtly practical artefacts: conventional institutions, where a collective activity —say a classical auction— is run according to some institutional rules. We can simply look into the principles of how such conventional institutions work and take them online The following is an informal characterisation of online institutions as a multiagent system and its distinguishing features are discussed below. A more rigorous characterisation is in [4].[5]

[5] In OIs, like in any multiagent system, one can identify two primitive components: the active agents in the institution and the environment that enables and governs the interactions of those agents. In OIs, the environment itself includes a limited ontology (i.e., a set of entities that are involved in the description of the facts that may at some point hold in the institution, *affordable actions and feasible events*) that is common to all the active agents. Because we mean to capture the governance functions of conventional institutions, the environment also provides the devices that determine whether agents can enter the environment, and provides the devices that govern the activity of agents (communication, display of information and norms). Note that it is possible that in a given OI there may be *institutional agents*. These agents would be `institutional" in the sense that their behaviour is the responsibility of the OI and, if they are artificial agents, their decision-making model is given by the OI.

Construct 1. **Online institutions** is the class of *multiagent systems* that are:

(i) *open*: there is an "inside" and outside" of the OI, and a priori one knows not which agents are active inside,
(ii) *hybrid*: human and software agents[6]
(iii) *situated*: it is part of the actual world and functions within a particular socio-technical context
(iv) *online*: the OI is a technological entity, and agents interact with it and among themselves via the environment(s)in which they are situated
(v) *regulated* (all agent interactions conform to some constraints declared and enforced by the OI)
(vi) *state-based* (the state of the institution is characterised at a given instant by a collection of institutional facts, namely permissions or prohibitions,[7] powers, obligations and any domain facts relevant to the function of the institution), see also construct 2, and
(vii) satisfy the *observability* and the *dialogical* stances, see construct 3 and 4 respectively. •

First, we want to refer to the way an OI is governing a collective interaction that evolves over time, so we define the **institutional state at time t** as the set of facts that hold in the institution at that point. This gives us a new feature:

Construct 2. OIs are **State-based** which means that the institutional state of the system is unique and the same for every participating agent, and that only afforded institutional actions and feasible institutional events can change it.•

Second, to be able to assess that the state changes according to the conventions established by OI, we adopt

Construct 3. the **observability stance.** At any point in time, the (institutional) state of the world is a finite set of observable facts.•

Finally, in order to enforce this constraint, we restrict feature (iv) above to OI interpretations of agent actions and hence enforce institutional conventions

Construct 4. **Dialogical stance.** All institutional interactions are *illocutory acts* that are mediated by the OI *interface*.•

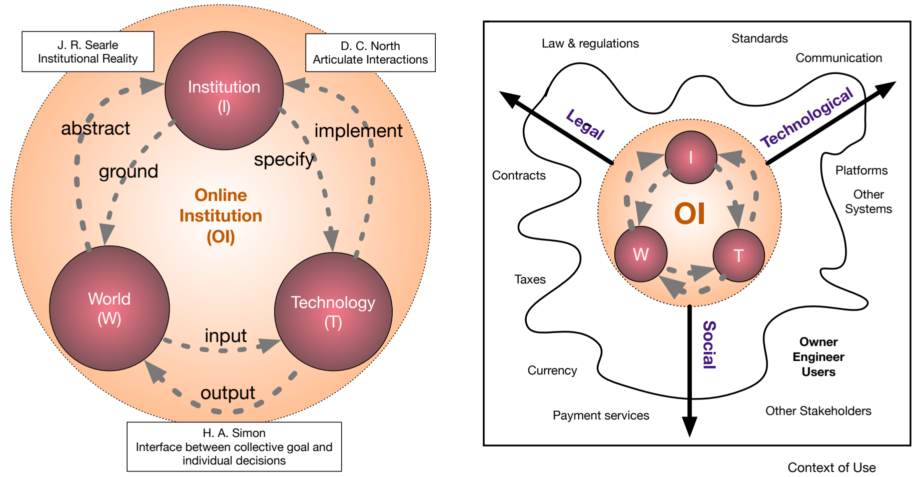
3 An abstract view of OIs: the WIT model

We can use the *WIT model* represented in Fig. 1 to characterise an OI as the combination of three components:

[6] It is not necessary that humans are involved in every OI. What is assumed is that the decision-making of participating agents is "opaque" or not accessible by the institution. The point of this property is to contemplate the possibility that agents may be heterogeneous, hence incompetent or malevolent, and belong to different principals.

[7] Permissions implies a prohibitive institutional environment in which all actions are prohibited unless explicitly permitted, where prohibitions imply a permissive environment in which all actions are permitted unless explicitly prohibited

- $\mathcal{W}$ that corresponds the fragment of the real world that is *relevant* for the activity that is performed within the OI,
- $\mathcal{I}$ is an abstract representation of $\mathcal{W}$ that establishes the "rules of the game" and thus provides the specification of how the OI is meant to operate, and
- $\mathcal{T}$ that corresponds to the information technology that implements and supports it.



(a) Relations between the $\mathcal{WIT}$components and their relationship with three conventional views on institutions (Searle [13], North [8] and Simon [15])

(b) The Situated OI with its three compatibility requirements

Fig. 1: The "stand-alone" and "situated" views of an Online Institution (from [6])

Figure 1 suggests how the three components are interrelated and how these relationships reflect the classical views on intuitions. Our use of the term "institution" and our characterisation of OI purposely reflect four conventional interpretations of the term as Fig. 1 shows. Searle's proposal of a creation of an institutional reality that is different than the *crude* reality, see[13] (Features (i) and (vi) above; and the relationship between the $\mathcal{W}$ and $\mathcal{I}$ below); North's understanding of institutions as artificial constraints that determine the "rules of the game" [8] (Feature (v)); and Simon's view of institutions as interfaces between individual decision-making and a collective objective [16] (Feature (iv), the dialogical stance and the relationships between the $\mathcal{W}$ and $\mathcal{T}$ views below). Ostrom's ADICO framework and her outlook on the social insertion of institutions [9] are addressed in the expressiveness of the $\mathcal{I}$ view of OIs and of a situated institution (Feature (iii) and below).[8]

[8] Notice that while there is a clear distinction between the institution itself and the participating agents, there may be some participating *institutional* agents that are designed as part of the institution, which is responsible for their decision-making.

The OI concept has evolved over the years in the MAS community over a long period, as various mechanisms for social coordination have been proposed (see e.g., [1]). We have in previous publications referred to OIs as socio-cognitive technical systems or a hybrid online social systems (see [17,3,17,4,7])

3.1 A stand-alone OI

The OI is the combination of the three $\mathcal{WIT}$ components but it is convenient to look at them separately because they make explicit different features that need to be articulated in order to have a well-defined working OI. The three parts work together to ensure that an agent action can be interpreted against the "rules of the game" and thus correctly affect the relevant part of the world. This intuitive description may be made precise by claiming that an OI ought to be "cohesive".

Cohesiveness is based on the postulate that OIs are state-based and that only some actions and events can change the state of the institution (see Feature (vi) and Feature (vii) in Cons. 1). Technically speaking, the property assumes that (i) the (crude) agents, actions and events in $\mathcal{W}$ correspond to agent Identifiers, abstract actions, and events in $\mathcal{I}$ and to agent processes and inputs in $\mathcal{T}$; and (ii) that there is a "state of the institution" that is defined by states that are specific for each view ($\mathcal{W}$, $\mathcal{I}$ and $\mathcal{T}$). Thus, cohesiveness means that if at a given time the state of $\mathcal{W}$ changes (because a crude event or action is deemed "institutional"), the state of $\mathcal{I}$ and the state of $\mathcal{T}$ change accordingly. That is,

Property 1. **Cohesiveness** An OI is cohesive if the three views are isomorphic with respect to actions and events.●

The OI is an entity that is meant to be part of the world but, independently of the particular context where it will be situated (in Cons. 1, Feature (iii)), it needs to be —on its own— a distinct working entity (Cons. 1, Features (i) and (iv)) and its contents not contaminated or altered by the external world. Thus it should satisfy the following:

Property 2. **Integrity.** An OI is integral if (i) only those agents that are admitted by the OI are afforded an interface; (ii) the interfaces work correctly (i.e., only admissible institutional inputs enter the OI and only institutional outputs leave); (iii) institutional data is incorruptible (communication works, inputs are processed correctly, results of processes are persistent and outputs are properly sent); and (iv) the OI is impervious (only that information that is requested, admitted or emitted by the OI enters or leaves the OI).●

3.2 A situated OI

By definition (Feature (iii) in Cons. 1), OIs are meant to support interactions that will have an effect in the real world, and actual individuals and organisations are involved in its operation. Therefore, in particular, to be effective they have to be *compatible* with the real world along three main areas of concern: those aspects of the actual world that (i) allow its online operation (technological standards, communication infrastructure, data, IP devices, ...); (ii) those aspects that validate and make the transactions legally effective (contracts, applicable regulations and law), and (iii) those social aspects that are relevant for its successful operation (economic conditions, social norms, commercial and working practices,...).

Property 3. **Compatibility** An OI needs to comply with *technological, legal and socio-economic* standards, practices, and norms that enable its effective operation in the environment where it is embedded.•

There is a second property that situated OIs should have. It corresponds to the realisation that no matter what the actual purpose or functionalities of the OI, and in addition to any other direct or indirect stakeholders of the OI, there are *three stakeholders* that are always involved in the *design, construction and deployment* of the situated OI: the eventual users of the OI, the team of engineers, designers, and support people that are in charge of the construction and maintenance of the OI and the owner, (the entity) who commissions, releases and exploits the OI.

Property 4. **Design Stakeholders** Any OI always has three types of *design stakeholders*: owner, builder and users.•

3.3 The $\mathcal{WIT}$ design pattern

The $\mathcal{WIT}$ Model serves as the blueprint for the design of OIs, in the sense of Alexander's "design patterns" [2]. The four salient elements have already been mentioned: the separation of concerns into the six arrows that link the $\mathcal{WIT}$ views: abstraction/grounding, specification/implementation, and input/output; the existence of three essential design stakeholder types (user, builder and owner); the two stand-alone OI properties: cohesiveness and integrity; and the three types of compatibility requirements of the situated OI (legal, technological and socio-economic).

4 OI-based assumptions for Conscientious Design

We are interested in a restricted version of the *Value Alignment Problem* that applies to the design and building of OIs, not AIS in general. By restricting our focus to OIs we are able to demonstrate what is needed to get to our goals of engineering systems which are *provably aligned with human values.* The reason for choosing OIs to characterise a restricted form of the VAP is because OIs justify some assumptions that facilitate the engineering of values. The following is an attempt to make those assumptions explicit and to illustrate how these assumptions are put to work.

The conventional understanding of values. We assume a rather standard motivational/cognitive view of values (compatible with e.g., Schwartz [12]) with the following properties:

V. 1 Values motivate goals.
V. 2 Values justify actions.

V. 3 Values serve as criteria to determine preferences between states of the world.
V. 4 Values legitimise goals
V. 5 In the assessment of a state of the world or in justifying an action, several values may simultaneously apply and these may be in conflict.
V. 6 Value imbuing and assessment are contextual.●

Assumptions for the Value Alignment Problem. There are three implicit assumptions in the wording of the Value Alignment Problem that decompose it into three subproblems: (i) that one can choose some values that the system should comply with, (ii) that those values can be embedded into the system, and (iii) that one can objectively assess the alignment of the system with those values.

*Vap.*1 The VAP can be decomposed into two problems: value imbuing and the assessment of value alignment.
*Vap.*2 One needs to be explicit about the values that will be embedded in a given OI, and determine the alignment of the system with respect to all those values (see [10]).
*Vap.*3 We understand that "provably aligned" is meant as an informal but objective (not necessarily proof-theoretic) way of determining that an AIS is aligned with a value or a set of values.●

Assumptions for the Value Alignment Problem in OI. Because we are concerned with the VAP only with respect to OIs, we make explicit the way that the VAP is interpreted in OIs with the following additional assumptions:

VapOI. 1 In OIs, the VAP concerns the engineering of values in two different entities: in the governance of the multiagent system, and in the decision-making model of individual (artificial) *institutional agents*.[9]
VapOI. 2 We assume that the process of engineering values in an OI can be organised in a cycle with three main stages whose outcome is the specification (in $\mathcal{I}$) of how values will be implemented in the OI (in $\mathcal{T}$).
i *Contextualisation* in OIs: The choice of values depends on the domain of application of the OI, the stakeholders and the separate design concerns and compatibility requirements of the OI (as induced by the WIT-design pattern). We assume that such contextualisation applies also to the imbuing and assessment decisions.
ii *Imbuing* can be split in two that are closely linked with assessment: (i) *interpretation* (the features that make the value observable and its alignment objective) and (ii) *instrumentation*. the means that conditions the outcomes of actions accordingly. In OIs this is part of $\mathcal{I}$.
iii *Assessment.* How to determine whether an OI is "provably" aligned with a vale and with a set of values.[10]●

[9] This engineering process is applicable to autonomous artificial agents in general, as well as OIs, but in this paper, we do not address the specifics of ethical reasoning in autonomous agents.

[10] As we discuss below, assumes that values can be protected or fulfilled to a certain degree that can be measured and that the protection or fulfilment of a combination of values can also be measured.

The Objective Stance . This fundamental assumption makes explicit how to interpret "provability" of alignment (*VAP.*(3)) in the case of OIs and motivates the working assumptions needed to eventually engineer values in OIs

OS. 1 **[Objective Stance:]** The alignment of an OI with a value can be measured as a function of the state of the world.•

The rationale is as follows: First, by definition, OIs are *state-based* (Cnstr. 2). Second, by definition the (*Observability stance* (Cnstr. 3), the institutional state is a *finite set of observable facts*. Third, from *Val.*(3), we assume that values can determine preferences over the state of the world, and therefore, there are some states that are preferable for a given value. Fourth, these preferred states can be defined as *goals* that are motivated for that value (and possibly other values) (*Val.*(1)) and legitimised by it (and possibly other values) (*Val.*(4)). Therefore, goals are representable as a function of a finite set of observable facts and a value corresponds to a composition of all its goals.

In order to make this *Objective Stance* fully operational we still need to make explicit four additional *assessment assumptions*:

OSa. 1 *Goal satisfaction function:* Given a goal for a value v, one can define a function that, for each state of the world, measures the degree of satisfaction of that goal (with respect to the value) in that state.

OSa. 2 *Goal aggregation function* Given a value and the set of all its goals, one can define a goal aggregation function that, for each state of the world, measures the degree of satisfaction of the value as a combination of the satisfaction of its goals, in that state.

OSa. 3 *Value aggregation function:* Given a set of values with their own goal aggregation functions, one can define a value aggregation function that determines, for each state of the world, the degree of satisfaction of the combination of all the values, in that state.

OSa. 4 *Value alignment assessment:* Based on the above one can define functions that capture the different interpretations of alignment with respect to particular value interpretations. •

Assumptions about instrumentation. Since the state of the world changes when an action takes place (Construct(3)), an action can contribute towards a given goal. Moreover, since only events and afforded actions can change the state of the world, the way to imbue values in an OI is to enable, curtail, promote or discourage individual actions or to modulate events in order to achieve the intended goals. Consequently, one chooses to incorporate an instrument in the governance of an OI because that instrument has positive effects on the goals associated with the value.

On the other hand, actions can be seen as functions from the state of the institution into the state of the institution. Moreover, since the institution's state is finite, one can measure the effects of an action with respect to a goal through the changes the action produces in the facts associated with the goal. Consequently, any given action can have measurable effects towards any goal, and one can ascertain trade-offs in the effects of any particular action with respect to the different goals and values.

In other words:

*Ins.*1 Let G be goal whose observable facts is set F; then, for each action α that affects F one can measure the effect of α towards G by the change of the degree of satisfaction of goal G and likewise for other goals.

*Ins.*2 For each goal G one can choose instruments that either promote actions that have positive effects on G, or discourage actions that may have a detrimental effect.ç

*Ins.*3 There are three types of value-imbuing *instruments* for OIs: (i) *affordances* (enable agents to take some actions (recognising (crude) actions in $\mathcal{W}$) that can have an institutional effect, (ii) *norms* (in $\mathcal{I}$) that regulate the conditions and effects of institutional actions, and (iii) *information* that may influence the decision-making process of policy subjects (made available to participating agents through $\mathcal{T}$). •

Assumptions on the use of the WIT design pattern and Conscientious Design value categories. Finally, the $\mathcal{WIT}$ pattern provides assumptions for heuristics on value contextualisation and assessment features; and Conscientious Design Values, for heuristics to identify and tailor goals, and to define value alignment criteria. Other design assumptions about conscientious design, —not $\mathcal{WIT}$–specific— were included in [6].

*WIT.*1 Values and their engineering should be contextualised for the OI domain (i.e, the purpose of the OI, taking into account the $\mathcal{W}$ ontology, afforded actions, and roles of participating agents), the three design stakeholders (user, owner, builder), the integrity and compatibility properties of the OI and the six separate design concerns ($\mathcal{WIT}$ "arrows": abstraction, grounding, specification, implementation, input and output).

*WIT.*2 Conscientious design values (thoroughness, mindfulness and responsibility) can be used in the functions to ascertain the global alignment of the OI and to identify goals in the WIT contextualisation process. •

5 Example heuristics for value engineering OIs

The following remarks illustrate how the assumptions we made explicitly may be used to design value-aligned OIs.[11]

An OI is built with some general purpose in mind, that needs to be properly contextualised and interpreted (*VapOI.* 2, *WIT.*1 and *WIT.*2). Values inform the way this purpose is achieved: they clarify goals, measure and compare the outcomes of actions, and determine what governance instruments provide the best alignment (*OS*). More specifically:

Heuristic 1. Values enable *courses of action*. •

In practice, this simply means that

(i) Values serve to adopt explicit goals, these goals need to be made precise enough (*OS*) so that they reflect the needs and motivations of each and all stakeholders and of the different design concerns (*WIT.*1). Values consequently clarify and validate the ontology that needs to be incorporated into the OI (affordances, facts and, in summary, the relevant fragment of reality).

[11] These heuristics complement the ones in [6].

(ii) Goals are validated by values: each goal is a desirable state of the world for some value and the governance instruments lead actions toward that state (*Ins.*3). This happens for every goal of every value,

(iii) Value aggregation functions allow the assessment and comparison of those classes (*OSa.* 2).

(iv) The way that values are imbued in the OI, —as governance instruments that condition the evolution of the institutional sate— modulates the activity of participating agents towards desired end-states (*Ins.*2); that is, values define the space of interaction.

(v) The assessment of value alignment clarifies the preferable courses of action; because it measures the consensual satisfaction (of the consensual goals and values, for all contextualised values), the satisfaction for each stakeholder and the relative cost/benefit of alternative governance instruments (for the consensual values).

Heuristic 2. *Value contextualisation and imbuing.* OI values can be contextualised and imbued in four successive stages: (i) values for the application domain and CD categories for the consensual preferences of the three design stakeholders towards the OI , (ii) for the individual preferences of each of the three design stakeholders of the OI; (ii) then for the compatibility requirements of the situated OI; and, finally, (iii) for the six WIT-articulation design concerns (abstraction, grounding, specification, implementation, input and output). •

Value *interpretation* (*VapOI* 2.1i)) is achieved by defining value-specific goals, and for each goal the features that are involved in the assessment of the contribution of that goal to that value; whereas *instrumentation* is achieved by identifying the means to achieve those goals (*Ins.*1,3). In turn, value assessment (*VapOI:* 2.iii) is achieved by adopting goal measurement functions, and goal and value aggregation functions (*OSa.* 1-3); as well as a way of assessing the impact (positive and negative effects) of the instruments with respect to all the goals (*Ins.* 2). Therefore while establishing "courses of action" requires consensus among all stakeholders, different stakeholders' preferences should still be considered in the final assessment of value alignment. We articulate these remarks with *OSa.* 1-4 in mind:

OI design should respond to the needs and interests of all stakeholders. Therefore, value choice, interpretation, instrumentation and assessment, are consensual. But individual stakeholders may hold different values and may also use their own satisfaction and aggregation functions to assess the consensual goals and instrument effects. The assessment of value alignment should reflect these considerations.

Heuristic 3. OI's values, goals, goal satisfaction functions, goal aggregation functions, value aggregation functions and value instrumentation are *consensual*. •

Heuristic 4. *Each stakeholder* holds their own values, goal satisfaction and aggregation functions, and value aggregation function. These apply to the achievement of OI's goals and the assessment of instrument effects and will therefore provide that stakeholder with the elements for their own assessment of value alignment. •

Recall that the aim of our proposal is to imbue values in an OI in such a way that the OI is *provably aligned* with them (*Vap.3*). Based on the previous two heuristics, we propose to address value alignment through a combination of three types of alignment that keep the consensual and individual differences in mind:

Heuristic 5. *Value alignment* can be assessed as a combination of three assessment procedures:

1. An assessment of the *effectiveness* of the governance instruments to satisfy the OI goals and the resulting aggregated value satisfaction based on consensual features (values, goal satisfaction and aggregation function, and value aggregation functions (*Heur*. 3).
2. Assessing how *adequate* are the governance instruments for producing the alignment in terms of their direct and indirect effects (equally effective sets of instruments may have different cost-benefit trade-offs)
3. Assessing how *acceptable* are the governance instruments for the stakeholders. Acceptability combines the individual assessments of all the stakeholders. This individual assessment is the stakeholder's assessment of the effectiveness and adequacy of the instruments with respect to their own values (*Heur*. 4), not the OIs values. •

With the previous heuristic in mind, we now list heuristics that apply to the consensual aspects of the OI design: OI goals, governance instruments, goal satisfaction functions, goal aggregation functions and value aggregation functions.

Heuristic 6. *Choice of values and their goals* can be addressed as a goal decomposition process (which is accompanied by a means-ends analysis). The resulting tree (for each value) is rooted in an abstract "tellic" value and its leaves are consensual goals. •

Goals determine the facts that need to be observable and there should be a consensus on how to assess, for any state of the world the extent to which that goal is satisfied (*OSa*.1). There should also be a consensus on how the combined satisfaction of those goals amounts to a satisfaction of the value that motivates them (*OSa*.2). From *OS* we propose a pragmatic compromise for goal satisfaction and goal aggregation: (i) an *objective function* that defines an ordering of the states of the world with respect to how good that state is for the satisfaction of the goal, and (ii) a threshold *–aspiration level*— for each objective function that determines the minimal level of satisfaction for that goal. The same heuristics apply to value satisfaction (as a combination of goal satisfaction (*OSa*.2) and value aggregation (as a combination of values (*OSa*.3)). This way we can make goals (and values) minimally commensurable and thus obtain goal and value aggregation functions.

Heuristic 7. Goals determine an *objective* function that gives the degree of satisfaction of the goal for each state of the world (with respect to a value). For each goal, there is an *aspiration level* that determines the minimal value of a state that achieves the satisfaction of the value. Likewise, value satisfaction is an objective function for goal aggregation, with an aspiration level for value satisfaction. •

One can think of these objective functions for goals as a way of imposing a total order on the states of the world with respect to each goal, as a primitive sort of utility function of that goal, with positive and negative utilities separated by the aspiration level. Value satisfaction is determined by a composition of the goal satisfaction functions and amounts to an aggregated utility function of the combined satisfaction of its goals, with the value aspiration level as its threshold.

Notice that, as a side-effect, Heuristic 7 suggest how goal and value aggregation functions induce an ordering of goals and values.

Finally, in order to address the *Adequacy* and *Acceptability* assessments in Heur. 5 we propose, based on the Ins. assumptions, Heuristic 8 for identifying instruments that contribute to a policy goal and Heurisitc. 9 for determining the direct and indirect impact of an instrument and deciding to incorporate or not that instrument in the OI governance.

Heuristic 8. Values are imbued in the OI as instruments that modulate those affordances that affect the parameters of an OI goal. •

In practice, for each (consensual) goal, the process is first, based on *Ins*.1, to identify affordances that affect the observable parameters involved in the assessment of the goal and explore for each affordance the direct effects on that and other goals. Second, instrument the affordance (*Inst*.2) to achieve the best effects; that is, enable (add it as a new affordance if needed) or inhibit it (or discard it), regulate it (foster, discourage, curtail or prohibit), or design information to incline participating agents decisions to those effects (*Ins*.3).

Heuristic 8 alone would produce too many instruments. One way to navigate this problem is to do the instrumentation incrementally, by looking into the cost-benefit trade-offs of the instruments that may be more relevant for an effective alignment of the OI goals. Here, one can use the value aggregation functions to prioritise values and the goal aggregation functions to prioritise goals. Then combine these two orderings and pick the affordances that impact the most important goals and keep those that are the most adequate. The instrument thus operates on a subset of affordances and assesses their effects on the satisfaction of the most relevant goals.

Heuristic 9. Priorise values and their goals, an instrument first those affordances that affect most the more significant goals. Measure and compare the effects of instruments on the prioritised goals. •

6 Closing Remarks

1. There are three points about the theory of values the we leave for further discussion. First, our assumption of an objective stance establishes a correspondence between values and their consequences only as far as those values bear upon the specific OI. We justify this correspondence on pragmatic grounds —provability of alignment and agreement and validation of design requirements with stakeholders— and do not find it necessary to commit to full consequentalism. That is why we left out the problem of imbuing values in the decision-making model of artificial institutional agents out of this paper (see below). Second, our proposal of the *Objective Stance* (*OS*), suggests an analogy between values and organisational goals as discussed by Simon [14] that we outline in the discussion of Heuristic 1. The other heuristics are meant to qualify the extent of that analogy. Third, in accordance with our previous remark, heuristics 5 and 7 are

meant as a minimal way of making values commensurable, again supported by pragmatical requirements, and several forms of multiobjective decision-making schemes may well apply in different cases.

2. In this paper, we have mentioned Conscientious Design value categories tangentially. Our discussion of these categories in [7,6] provide substance and scope to the assumptions and heuristics we present here.

3. Our discussion in this paper has been centred on the governance provided by the OI and has mentioned ethical decision-making only in passing. As we mentioned in [6], one can engineer values in an artificial agent in three ways: reactive behaviour learned behaviour or actual value-driven reasoning. For this last way, the heuristics we propose here also apply to ethical reasoning.

4. Our heuristics for value engineering may also apply to other artificial autonomous intelligent systems but must be revised for such extension. Nevertheless, the class of OIs is interesting on its own behalf for its intrinsic complexities but also because it encompasses an increasing large class of existing AI-enabled systems.

5. The Value Alignment Problem is only one instance of the relevance of values for AI in general. We like to think that our proposal, albeit centred on OIs, contributes to a wider project of an AI-driven theory of values.

7 Acknowledgements

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

References

1. Aldewereld, H., Boissier, O., Dignum, V., Noriega, P., Padget, J.: Introduction, pp. 3–9. Springer (2016). https://doi.org/10.1007/978-3-319-33570-4_1, http://dx.doi.org/10.1007/978-3-319-33570-4_1
2. Alexander, C.: A pattern language: towns, buildings, construction. OUP (1977)
3. Christiaanse, R., Ghose, A.K., Noriega, P., Singh, M.P.: Characterizing artificial socio-cognitive technical systems. In: Herzig, A., Lorini, E. (eds.) Proceedings of the European Conference on Social Intelligence (ECSI-2014). pp. 336–446. CeUR (2014), https://ceur-ws.org/Vol-1283/
4. 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))
5. Noriega, P., Verhagen, H., d'Inverno, M., Padget, J.A.: A Manifesto for Conscientious Design of Hybrid Online Social Systems. In: Cranefield, S., Mahmoud, S., Padget, J.A., Rocha, A.P. (eds.) COIN@AAMAS, Singapore, May 2016, COIN@ECAI, The Hague, The Netherlands, August 2016, Revised Selected Papers. LNCS, vol. 10315, pp. 60–78. Springer (2016)

6. 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)
7. Noriega, P., Verhagen, H., Padget, J., dInverno, M.: Ethical online AI systems through conscientious design. IEEE Internet Computing **25**(6), 58–64 (2021)
8. North, D.: Institutions, Institutional Change and Economic Performance. CUP (1991)
9. Ostrom, E.: Governing the Commons. The Evolutions of Institutions for Collective Action. Cambridge University Press, Cambridge (1990)
10. van de Poel, I.: Embedding values in artificial intelligence (AI) systems. Minds and Machines **30**(3), 385–409 (2020)
11. 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-ai#26015, [Online] Retrieved 12 december 2022
12. Schwartz, S.H.: An overview of the Schwartz theory of basic values. Online readings in Psychology and Culture **2**(1), 11 (2012)
13. Searle, J.R.: The Construction of Social Reality. Allen Lane, The Penguin Press (1995)
14. Simon, H.A.: On the concept of organizational goal. Administrative Science Quarterly **9**(1), 1–22 (1964), http://www.jstor.org/stable/2391519
15. Simon, H.A.: The Sciences of the Artificial. MIT Press, third edn. (1996)
16. 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)
17. Verhagen, H., Noriega, P., d'Inverno, M.: Towards a design framework for controlled hybrid social games. In: Social Coordination: Principles, Artefacts and Theories, SOCIAL.PATH 2013 - AISB Convention 2013. pp. 83–87 (04 2013)