Addressing the Value Alignment Problem through Online Institutions

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] Artificial Intelligence Research Institute (IIIA-CSIC), 08193 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. As artificial intelligence systems permeate society, it becomes clear that aligning the behaviour of these systems with the values of those involved and affected by them is needed. The value alignment problem is widely recognised yet needs addressing in a principled way. In this paper, we investigate how such a principled approach regarding online institutions — a class of multiagent systems — can inform us on how the general value alignment problem can be approached.

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 modulate artificial autonomy is to incorporate ethical considerations into the design and construction of artificial systems. In particular, to conceive a form of ethics as a means of controlling that autonomy. Stuart Russell articulated this intuition as the challenge *to build systems that are provably aligned with human values* and referred to it as "the Value Alignment Problem" (VAP) [15].

The Value Alignment Problem can be understood as an engineering challenge that needs a rigorous approximation to the notion of "value" if one intends to evidence the degree to which an AIS provably aligns with a set of values. We propose to address the VAP challenge with a principled approach that starts by circumscribing our treatment of the VAP to a particular class of AIS: *Online Institutions* (OI), then establish relevant conceptual distinctions for this scoped version of the problem and define constructs that capture those distinctions. With these elements – and the background of "conscientious design" [9,11] – we can then propose heuristics and methodological guidelines for the design, operation, and monitoring of OIs that are provably aligned with some values. Although this is a restricted version of the VAP, we claim that value alignment for OIs involves the same requirements as the full VAP, the salient distinction being that OIs

have, by definition, some particular features that allow a precise characterisation of value engineering.

This paper is an argument for this claim, organised in three parts. First, in order 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 assumptions about values predicated on the definition of an online institutions. Finally, in Sec. 5, we enumerate specific heuristics that illustrate how these (now explicit) assumptions support the dual empirical problem of embedding values in a system and assessing that the resulting system is provably aligned. The final section elaborates on our assumptions and gives some context for future work.

This paper is another step towards our goal of designing a principled approach to the VAP. The key technical details and their contextualisation that complement the argument we present here can be found in four previous publications. (i) In "A Manifesto for Conscientious Design" [9] we outlined a research programme for value-driven design of artificial intelligent system; (ii) "Anchoring Online Institutions" [8] contains a more systematic presentation of the contents of sections 2 and 3; (iii) In "Ethical online AI systems through conscientious design." [11] we outlined 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" [10] we discussed the value operationalisation process and some heuristics for how to attack the process.

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. One can simply look into the principles of how such conventional institutions work and translate them online. As we discuss next, online institutions interpret that intuition in a way that is convenient for all sorts of applications. Several commercial systems fall into the class of OIs, for instance, *Uber* and *Amazon*, and in [10] we use an ideal online ticketing service as a typical OI (to illustrate the value engineering process).

The following is an informal characterisation of online institutions as a multiagent system, and its distinguishing features are discussed below in Sec. 3. A more rigorous characterisation is in provided [8].[5]

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

[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 ([4,7]) . In OIs, the environment itself includes a limited ontology – which includes a set of entities that are involved in the description of the facts that may at some point hold in the institution, as well as enabling 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, as well as the devices that govern the activity of agents (communication, display of information, enforcement of institutional constraints).

(i) *open*: there is an "inside" and an "outside" of the OI, and while participants may enter and leave the OI, *a priori* one knows not which agents are active inside the OI;
(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 are subject to some constraints that are declared and enforced by the OI;[7]
(vi) *state-based*: the institutional state is unique and the same for every participating agent, and only enabled institutional actions and feasible institutional events can change it;
(vii) satisfy the *dialogical* and the *observability* stances (see constructs 3 and 4 respectively). •

Features (vi) and (vii) are included in this definition because the OI is governing a collective interaction that evolves over time. Thus, we need to refer to an institutional state that changes, but changes when and only when *institutionally recognised* events and actions take place (and this last part is supported by Features (iv), (v) and constructs 3 and 4).

Construct 2. **The institutional state at time** t (s_t) is the set of facts that hold in the institution at time t.[8]•

The Dialogical Stance supports the enforcement of Feature (v) above (by filtering all potential potential changes through the interface implicit in Feature (iv)). The Observability Stance allows us to detect that a change has taken place.

Construct 3. **Dialogical Stance.** All institutional interactions are *illocutory acts* that are mediated by the OI *interface*.•

Construct 4. **Observability stance.** At any point in time, the institutional state of the world is a finite set of observable facts.•

[6] Humans don't need to be involved in every OI; what is in fact assumed is that the decision-making of participating (non-institutional) agents is "opaque" or not accessible to the institution. The point of this property is to acknowledge the need to govern the behaviour of participating agents that may be heterogeneous, incompetent, malevolent, or belong to different principals.

[7] This feature may be realised in different ways; one is to think of OIs as *normative multiagent systems* (see [3]); however, in a given OI, the particular representation of institutional constraints and their enforcement is reflected in the institutional model (Ψ of $\mathcal{I}$) see Sec. 3.

[8] We can be more precise defining it as a point in the institutional space at time t. That is, $s_t \in \mathcal{S}_t = \times_{i=1}^{n} D_i$, where each D_i is a "domain", there is an initial state $\mathcal{S}_0$ that changes only when an event or an action performed by a participating agent complies with the active institutional constraints (actions and events are partial functions on $\mathcal{S}$)

The OI concept has been evolving over the years within the MAS community where various frameworks for social coordination have been proposed (see for example, [1]).[9]

3 An abstract view of OIs: the WIT model

In general terms, an OI establishes, enforces, and processes *capabilities and constraints* to govern a collective activity. To make this view concrete, we can use the *WIT model* represented in Fig. 1 to characterise an OI as the combination of three components:

- $\mathcal{W}$ corresponds to 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
- $\mathcal{T}$ consists of the information technology that implements and supports the OI.

In coarse terms, $\mathcal{W}$ is the working system that humans or their software counterparts interact with. Those interactions involve tangible objects and have effects on the perceivable physical reality. By construction, the OI determines in $\mathcal{W}$ the entities that are involved in those facts that are recognised in the state of the world (Construct 2). Also by construction, the OI will recognise that only certain actions and events can change it (Construct 1 (vi)) and since all recognisable actions are illocutionary actions mediated by the OI (Construct 1 (iv) and Construct 3), the OI has to enable those actions for participating agents through the interface in $\mathcal{W}$.

While $\mathcal{W}$ enables interactions in the real world, $\mathcal{I}$ establishes how those interactions acquire an institutional status. Intuitively, we say that an action is enabled if that action can be executed by an agent inside the OI and all the real-world conditions for its execution can be met and their effects can be acknowledged by the OI (we say that all the "physical constraints" can be met). However, to be deemed *institutional* (Construct 1 (v, vi)), an attempted action not only has to be enabled, it also needs to satisfy the artificial constraints that are the "rules of the game" that the OI imposes on participating agents. Correspondingly, $\mathcal{I}$ contains two models: an abstract model (Φ) of how that part of the world that is relevant for the OI functions (including the natural or physical constraints of the relevant part of the world); and another model (Ψ) that contains the artificial (institutional) constraints that govern those interactions (see [7] for a discussion of a metamodel to represent these constraints and their enforcement). In contrast with the entities of the real world that are part of $\mathcal{W}$, in $\mathcal{I}$ there are agent identifiers and constants and variables that stand for real-world entities and facts, and functions that stand for events and enabled actions that happen in the real world.

Finally, $\mathcal{T}$ includes data structures that model the state of the world, processes that correspond to the activity of real agents, the code that implements the constraints established in $\mathcal{I}$ and the rest of the technological platform that supports the operation of the institution (see a discussion in [7]). Figure 1a suggests how the three components

[9] We have referred to OIs as socio-cognitive technical systems and as hybrid online social systems in previous publications (see [21,5,21,8,11]).

are interrelated and how these interrelations reflect some conventional notions about institutions.

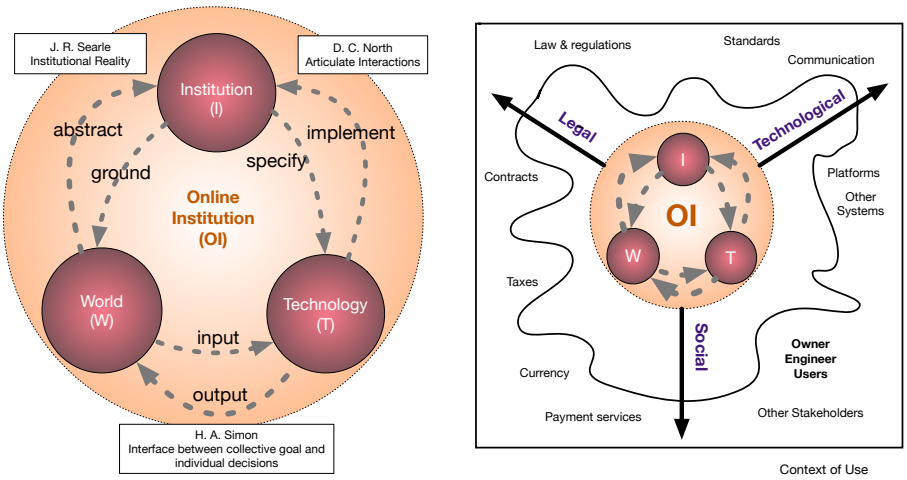
How the three components of the $\mathcal{WIT}$ model complement each other is established in Property 1 (cohesiveness) (see [8]). On one hand, the OI defines an "ontology". It determines in $\mathcal{W}$ what the relevant part of the world is and in particular establishes as part of that ontology what events and what actions of the real world are relevant in the OI. It also includes as part of that ontology all other entities of the real world that are needed for those events and actions to accomplish their intended institutional effects. In other words, we say that the OI provides *capabilities* to participating agents by recognising the real-world objects, events and actions that enable participating agents to act within the institution. But by excluding some real-world objects, events and actions from the *relevant* part of the world the OI also establishes constraints on what actions can be attempted and what events can take place. On the other hand, by definition, OIs are regulated multiagent systems (Construct 1, Feature (v)) that establish and – thanks to Feature (iv) in Construct 1 and the Dialogical Stance (Construct 3) – enforce the "artificial constraints" (beyond the physical constraints) that govern the empowered actions and are modelled on (Ψ).

Our use of the term "institution" and our characterisation of OI purposely reflect four conventional interpretations of the term. (i) Searle's distinction of an *institutional reality* that is different than the *crude* reality, (see [17]) is captured both in Features (i), (iii), and (vi) of Construct 1 and in the relationship between the $\mathcal{W}$ and $\mathcal{I}$ components of the $\mathcal{WIT}$ model in Fig. 1). (ii) North's understanding of institutions as artificial constraints that determine the *rules of the game* [12] is the reason for Feature (v), and become a specification in $\mathcal{I}$ that is implemented in $\mathcal{T}$), and (iii) Construct 1 captures Simon's view of institutions as *interfaces* between individual decision-making and a collective objective ([19]) through Feature (iv) and the *Dialogical Stance* (Construct 3). This view is reflected in the relationships between the $\mathcal{W}$ and $\mathcal{T}$ views in Fig. 1. Finally, (iv) Ostrom's ADICO framework and her outlook on the social insertion of institutions [13] are addressed, the first one, in the expressiveness of the $\mathcal{I}$ view of OIs (the way in which the artificial constraints are specified and enforced in Ψ) and, the second one, in the compatibility of a situated institution (Feature (iii) and the compatibility property discussed in the next section).

3.1 Properties of an OI

It is convenient to distinguish between the perspective of an OI as an entity on its own – a "stand-alone OI" (Fig. 1a) – as opposed to the perspective of an OI when we refer to it as an entity that is situated in its operating environment (Fig. 1b). In this section, we discuss two properties of the first perspective (Cohesiveness and Integrity) and one property of the second (Compatibility).

As noted above, the (stand-alone) OI is the combination of the three $\mathcal{WIT}$ components. 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. In fact, this decomposition becomes essential for the purpose of engineering values in an OI. In particular, the six arrows that connect the three components (Fig. 1a) are key for separating design concerns and the contextualisation of values (see below and [10]).



(a) Relations between the $\mathcal{WIT}$ components and their relationship with three conventional views on institutions (Searle [17], North [12] and Simon [18])

(b) The Situated OI with its three compatibility requirements

Fig. 1: The "stand-alone" and "situated" views of an Online Institution (from [10])

However, the three parts need to work in a cohesive way to ensure that an agent action can be properly accepted and executed following the "rules of the game" and thus correctly affect the relevant part of the world (see [8]).

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

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 (Features (vi) and (vii) in Construct 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 to 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.

In spite of being situated in a particular context (Feature (iii)), a stand-alone OI is itself, an entity whose functioning and contents should not be contaminated, exploited, or altered by the external world.

Property 2. **Integrity.** An OI is *integral* if (i) only those agents that are admitted by the OI are provided 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 information that is requested, admitted or emitted by the OI enters or leaves the OI).•

Finally, by definition (Feature (iii) in Construct 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 dimensions: those aspects of the actual world that (i) enable its online operation (technological standards, communication infrastructure, data, IP devices, ...); (ii) validate and make the transactions legally effective (contracts, applicable regulations and law), and (iii) are relevant for its successful social 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 wherein it is situated.•

3.2 Three remarks on Conscientious Design

In the introduction, we proposed to understand VAP as a design problem. In the next two sections, we make reference to ideas that contribute to "conscientious design" (CD), as formulated in [11,9]; here, we only touch upon three issues that support the design of OIs in which values are embedded.

Issue 1 At the core of CD is the understanding of design as a participatory process where the design stakeholders are involved in a cycle from the conception of the OI to its final decommissioning. This understanding assumes that values are taken into account in all the stages of the cycle, and that design stakeholders reach *consensus* at the different stages of the cycle (hence Assumption *CD*.1 in Sec. 4.6). This understanding also leads 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 at least *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, maintenance and operation, and monitoring of the OI and the owner, (the entity) who commissions, releases and operates and monitors the OI. Hence,

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

Issue 2 The $\mathcal{WIT}$ Model serves as the blueprint for the design of OIs, in the sense of Alexander's "design patterns" [2]. Its 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).

Issue 3 there are three CD value categories: *thoroughness*, *mindfulness*, and *responsibility* that encompass other value categories proposed for embedding values in AIS (a comparison with the values proposed in EU [6] and IEEE [20] is detailed in [11]). In particular, for this paper, these three categories serve to validate the contextualisation of values (*Heuristic* 2) and legitimise the assessment procedures proposed for value alignment (*Heuristic* 5).

4 OI-based assumptions for Conscientious Design

As stated in Sec. 1, we are interested in a version of the *Value Alignment Problem* that applies to the design and building of OIs, not of AIS in general. The reason for choosing OIs to characterise a version of the VAP is because OIs justify some assumptions that in turn facilitate value engineering. The following is an attempt to make those assumptions explicit and to illustrate how these assumptions are put to work.

4.1 The conventional understanding of values

We assume a rather standard motivational/cognitive view of values (compatible with e.g., Schwartz [16]) with the following properties:

V. 1 Values motivate goals.
V. 2 Values justify actions.
V. 3 Values legitimise goals.
V. 4 Values serve as criteria to determine preferences between states of the world.
V. 5 Values are contextual.
V. 6 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.•

4.2 Assumptions for the Value Alignment Problem

There are three implicit assumptions in the wording of the Value Alignment Problem that clarify three issues: (i) that one can choose some values that the system should support, (ii) that those values can be embedded into the system, and (iii) that one can assess the alignment of the system with those values.

*Vap.*1 The VAP can be decomposed into two problems: value embedding 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 [14]).
*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.•

4.3 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 for the design of 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*.

VapOI. 2 We claim 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 needs and preferences of design 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 embedding and assessment decisions.
- ii *embedding* can be split into two tasks 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 modulate 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 value and with a set of values.

4.4 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 specific 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.•

In other words, values can be represented as a function of a finite set of observable facts.[10]

In order to make this *Objective Stance* fully operational, however, we still need to make explicit three additional *assessment assumptions* that materialise the measurement

[10] The rationale is as follows: *First*, by definition, OI are *state-based* and by the (*Observability Stance* (Construct 4)), the institutional state is a *finite set of observable facts*. *Second*, from *Val.*4), we assume that values can determine preferences over the state of the world, and therefore, one can define a preference relation on the set of institutional states P_v for any given value v. *Third*, Since the state of the world is finite, one can choose preferable states for a given value v and define them as *goals* G_v that are motivated for that value (*Val.*1)) and also legitimised by it (*Val.*3)). *Fourth*, note that any goal (g) of value v_i will be included also in the preference relation (P_{v_j}) for every other value v_j (because g is one state of the world and because of *V.*6, several values may be involved in the assessment of a state of the world), however, it might not be a goal for v_j (g may or may not be in G_{v_j}.)

of the alignment of single values – identifying the degree of alignment of a value with a combination of the degree to which goals for that value are achieved – and also to deal with the alignment of multiple values simultaneously.

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

We label these assumptions "operational" because they need to be complemented with specific heuristics such as the ones we propose in Section 5. Such specific heuristics reflect different meta-ethical positions about values to some extent. [11]

Likewise, Assumption *OSa.*3 makes operational alternative notions of "objectively aligned" because it allows different ways of understanding the combination of several values (from Assumption *V.*6).[12]

4.5 Assumptions about instrumentation

Actions and events can be seen as functions that map the current institutional state into a new institutional state. Hence, since only institutionally acknowledged events and actions can change the state of the world, the way to embed values in the governance of an OI (in $\mathcal{I}$) is to enable, curtail, promote, or discourage individual actions or to modulate events in order to better achieve the intended goals. Analogously, institutional agents will have a value-aligned behaviour if and when their actions lead to the achievement of the intended goals. This alignment will depend either on predetermined behaviour that guarantees alignment with respect to specific goals by default, or because as institutional agents they are bound to comply with the institutional constraints and therefore the previous remark applies to their goal-driven reasoning.

Since institutional actions change the state of the institution, one can measure the effects (positive or negative) of an action α with respect to a goal g using the goal satisfaction function introduced in *OSa*.1. Note, though, that any given action can have measurable effects (positive or negative) towards the achievement of other goals and one can ascertain trade-offs in the effects of any particular action with respect to each one of the different goals and, ultimately, all values, using the satisfaction functions introduced in *OSa*.1 and *OSa*.2. In other words:

[11] For example. the conjunction of heuristics 2, 3 and 7 amounts to a weak form of consequentalism in which values are identified with goals but only for one specific OI and by the consensus of the design stakeholders who agree on the *consequences* of values.

[12] The heuristics we propose in Section 5 (notably *Heuristic* 5) are meant to allow value alignments that reflect the individual perspectives of the different design stakeholders, the consensual perspective and a combination of the two.

*Ins.*1 Let G be a goal whose observable facts is set F; then, for each action α that affects a fact $f \in F$, one can measure the effect of α towards G by the change of the degree of satisfaction of goal G; and likewise for any other goal G' whose observable facts include f.

*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.

*Ins.*3 There are three types of value-embedding *instruments* for OIs: (i) *actions* that are recognised (in $\mathcal{W}$) by an institution for a given agent to have an institutional effect; (ii) *norms* (in $\Psi \in \mathbf{I}$) that regulate the conditions and effects of institutional actions; and (iii) *information* that may influence the decision-making process of participating agents.[13] •

4.6 Assumptions from Conscientious Design

We make explicit three design assumptions that make the previous assumptions on values applicable in OIs. They are based on the remarks we made in Sec.3.2. The $\mathcal{WIT}$ pattern provides assumptions for heuristics on value contextualisation and assessment features, on Conscientious Design Value categories, for heuristics to identify and tailor goals, and to define value alignment criteria. Other design assumptions about design – not CD-specific – are considered in [10].

*CD.*1 Design stakeholders can reach *consensus* about OI values and goals, their satisfaction and aggregation, about the impact of instruments and about criteria for measuring alignment.

*CD.*2 Values and their engineering should be *contextualised* for (i) the OI domain (i.e, the purpose of the OI, taking into account the $\mathcal{W}$ ontology, enabled actions, and roles of participating agents); (ii) the three design stakeholders (user, owner, builder); (iii) the integrity and compatibility properties of the OI; and (iv) the six $\mathcal{WIT}$ separate design contexts (the six "arrows" of the $\mathcal{WIT}$ diagram: abstraction, grounding, specification, implementation, input, and output).

*CD.*3 Conscientious design value categories (thoroughness, mindfulness, and responsibility) can be used to ascertain *completeness and correctness* of goals in the WIT contextualisation process and in the functions to *ascertain* the global alignment of the OI. •

5 Example heuristics for *value engineering* OIs

The following remarks illustrate how the assumptions we made explicit above may be used to design value-aligned OIs.[14] An OI is built with some general purpose in mind that needs to be properly contextualised and interpreted (*VapOI.* 2 (Section 4.3), *CD.*1 and *CD.*2 (Section 4.6)). Values inform the way this purpose is achieved: they clarify

[13] In fact, one may implement institutional agents whose behaviour operationalise those three types of instruments values. For example institutional agents that perform discretionary norm-enforcement functions.

[14] These heuristics complement the ones in [10].

goals, assess and compare the outcomes of actions, and determine what governance instruments provide the best alignment (*OS*). In sum, values underlie the identification of what is relevant in the world and what "courses of action" lead to desired states of the world. More specifically:

Heuristic 1. An OI defines a *context for interaction* that enables actions and the constraints that modulate them. Values enable *courses of action* within that context. •

In practice, this means that

(i) Values serve to identify and 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 (*CD*.1 (Section 4.6)). Values consequently clarify and validate the ontology that needs to be incorporated into the OI (as part of the relevant fragment of reality ($\mathcal{W}$) and its abstract representation, Φ in $\mathcal{I}$.
(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) The way that values are embedded in the OI – as capabilities and governance instruments that condition the evolution of the institutional state – validates the ontology and modulates the activity of participating agents towards desired end-states (*Ins.*2); that is, values refine the space of interaction and enable courses of action.
(iv) The assessment of value alignment clarifies the preferable courses of action; because it measures the consensual satisfaction (of the consensual OI goals and values, for all contextualised values), the satisfaction for each stakeholder and the relative cost/benefit of alternative governance instruments.

Heuristic 2. *Value contextualisation and embedding.* OI values can be contextualised and embedded 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 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 *value 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 and 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).

While establishing "courses of action" requires consensus among all stakeholders, different design stakeholders' preferences should still be considered in the final assessment of value alignment. We articulate these remarks with *OSa.* 1-3 in mind:

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

Heuristic 4. *Each stakeholder* holds its own values, goal satisfaction, goal aggregation functions, and value alignment assessment functions. These stakeholder-specific functions apply to the assessment of the consensual OI's goals, and therefore do not necessarily coincide with the consensual assessments. Likewise, these stakeholder-specific functions are used for identifying and measuring the effects of the embedded values and will therefore provide each stakeholder with the elements for its own assessments of value alignment. •

Recall that the aim of our proposal is to embed 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 goal 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* the governance instruments are 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 (consensual) 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, and goal 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 the Objective Stance, (Sec. 4.4) 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. This way we can limit contradictions and tensions between the goals of the stakeholders and thus obtain a goal aggregation function.

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. •

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 aggregation functions induce an ordering of goals.

Heuristic 8. Values are embedded in the OI as instruments that modulate what is actionable to affect the parameters of an OI goal. •

Heuristic 8 implements the instrumentation assumptions but makes reference to goal parameters in order to identify the direct and indirect impact of an instrument. This allows the identification of trade-offs of the different instruments in order to address the *Adequacy* and *Acceptability* assessments in Heuristic 5.

In practice, for each (consensual) goal, the process —based on *Ins*.3 (Sec. 4.5)— is first to identify those actions that affect the observable parameters involved in the assessment of the goal and explore for each action its (direct) effects on that goal and (the indirect effects) in other goals, based on *Ins*.1. Second, to instrument the action (*Inst*.2) to achieve the best effects; that is, (i) to enable the action (add it as a new action in $\mathcal{W}$), (ii) to inhibit the action (or eliminate it form $\mathcal{W}$), (iii) to regulate the action (foster, discourage, curtail or prohibit), or (iv) design information to incline participating agents decisions towards those effects.

However, *Heuristic* 8 alone would produce too many instruments. One way to navigate this problem is to execute 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. To achieve this, one can use the goal aggregation functions to prioritise goals to identify the actions that impact the most important goals, and instrument only the most adequate (i.e. the ones with best cost-benefit trade-offs).

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

6 Closing Remarks

Earlier publications on online institutions (OIs) provide substance and scope to the assumptions and heuristics we present here. In particular, in [10] we describe an online ticketing system to motivate and illustrate heuristics for value engineering, and in [8] we use *Uber* to motivate and exemplify the definitions and properties of the WIT pattern. While in this paper we have mentioned Conscientious Design value categories only tangentially, a more detailed investigation of their nature and their relationship with other value taxonomies can be found elsewhere [11]

Our discussion in this paper has been centred on the governance provided by any OI and has only mentioned ethical decision-making in passing. As we mentioned in earlier work [10], one can engineer values in an artificial agent in three ways: reactive behaviour, learned behaviour, or symbolic and explicit value-driven reasoning. This is particularly relevant when designing autonomous *institutional agents* who become active in an OI on behalf of the institution itself (such as performing some norm-enforcement functions, for example). A full discussion of this aspect is beyond the scope of this paper but the heuristics we propose in Sec. 5 also apply in prinicple to the engineering of values in autonomous agents.

Our understanding of the VAP —as expressed in the assumptions in sections 4.2 and 4.3— makes it a "design problem". We actually propose a methodological approach to the design and construction of particular systems that would be provably aligned with a set of values, not a general solution of the VAP. Because we see the VAP as a design problem —and because one may wish to account for issues related to bounded rationality— our Objective Stance (Sec. 4.4) does not commit to any specific form of assessing value alignment or value aggregation. That choice would result from a consensus of the stakeholders who are involved in the design of a particular system.

The heuristics we propose for value engineering may also apply to other artificial autonomous intelligent systems but we have yet to explore this. Nevertheless, the class of OIs is interesting of itself for its intrinsic complexities but also because it encompasses an increasingly large class of existing AI-enabled systems.

Finally, 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 on an AI-oriented theory of values. We are looking forward to further investigating these possibilities.

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. Andrighetto, G., Governatori, G., Noriega, P., van der Torre, L.W.N. (eds.): Normative Multi-Agent Systems, vol. 4. Dagstuhl Publishing (2013)
4. Argente, E., Boissier, O., Carrascosa, C., Fornara, N., Mcburney, P., Noriega, P., Ricci, A., Sabater-Mir, J., Schumacher, M.I., Tampitsikas, C., Taveter, K., Vizzari, G., Vouros, G.A.: The role of the environment in agreement technologies. Artificial Intelligence Review **39**, 21–38 (01/2013 2013)

5. 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/
6. High-Level Expert Group on AI (AI HLEG): Ethics guidelines for trustworthy AI (Apr 2019), https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
7. 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.) Coordination, Organizations, Institutions, and Norms in Agent Systems X, Lecture Notes in Computer Science, vol. 9372, pp. 164–181. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-25420-3_11, http://dx.doi.org/10.1007/978-3-319-25420-3_11
8. 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))
9. 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)
10. 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)
11. Noriega, P., Verhagen, H., Padget, J., dInverno, M.: Ethical online AI systems through conscientious design. IEEE Internet Computing **25**(6), 58–64 (2021)
12. North, D.: Institutions, Institutional Change and Economic Performance. CUP (1991)
13. Ostrom, E.: Governing the Commons. The Evolutions of Institutions for Collective Action. Cambridge University Press, Cambridge (1990)
14. van de Poel, I.: Embedding values in artificial intelligence (AI) systems. Minds and Machines **30**(3), 385–409 (2020)
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-ai#26015, [Online] Retrieved 12 december 2022
16. Schwartz, S.H.: An overview of the Schwartz theory of basic values. Online readings in Psychology and Culture **2**(1), 11 (2012)
17. Searle, J.R.: The Construction of Social Reality. Allen Lane, The Penguin Press (1995)
18. Simon, H.A.: The Sciences of the Artificial. MIT Press, third edn. (1996)
19. 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)
20. The IEEE Global Initiative on Ethics of Autonomous and Intelligent System: Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, first edition (2019), https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf
21. 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)

A Response to reviewers

1. Reviewer 1
   (a) How safe is it to assume that values can be aggregated and their combined degree of satisfaction computed? That sounds awfully like a utility, as classical economists postulate (with dubious backing)

   RESP: We agree with this. The terminology has throughout been adapted to give a less mechanical flavor (alignment assessment instead of aggregation) and we also mentioned in various places that the value alignment assessment will depend on stakeholder and value combinations. We specifically addressed this issue in Sec 4.4 and in a remark in Sec 6 qualifying as "operational" the three assumptions that instantiate the objective stance, and the need to adapt such operationalisation of goal assessment and value alignment to the specific circumstances that affect an OI under design. We acknowledge also the link with utilitarianism but shading it as "consequentalism modulo a consensus of design stakeholders for a particular OI". The size and gist of the paper dissuaded us from discussing more practical issues like the possibility or extent of design consensus.

   (b) I am not too clear on "values enable actions". Maybe they legitimise actions – if that's what you mean

   RESP: We have not been able to find this exact statement in the text. It may be that you mean: "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" - where it it not the values themselves but how they are implemented in the OI that make certain actions possible (or impossible - curtail). On the other hand we subscribe the usual views that values motivate and legitimise goals, and claim that CD value categories are used to validate that the three major concerns of thoroughness, mindfulness and responsibility are properly taken care in the value-engineering cycle through the heuristics for contextualisation, interpretation and assessment of values.

   (c) Also, the requirement that goals "reflect the needs and motivations of each and all stakeholders and of the different design concerns" seems too strong in practice

   RESP: The goals reflect the outcome of the design process in which the values have been taken into consideration. This is hopefully more clear in the text now with our discussion on alignment assessment. We agree that this statement was too strong and this part is now removed from the paper.

   (d) In "Value aggregation functions allow the assessment and comparison of those classes", what are "those classes"?

RESP: This item (Heuristic 1 - iii) has been removed from the paper as have toned down / changed the value aggregation items.

(e) I see statements that "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" and that "choice of values depends on the domain of application of the OI". It would help to explain how you relate contextualization and values in a non-contextual abstraction of the online institution.

RESP: The text for OI property 1 has been changed to "In spite of being situated in a particular context (Feature (iii)), a stand-alone OI is, itself, an entity whose functioning and contents should not be contaminated, exploited, or altered by the external world" which we hope is clearer.

(f) It is clear that the authors are experts in this area. However, they could make a stronger effort to engage with the AAMAS literature. About the only citations I found that are not self-citations are to works by famous scholars (some of them – Simon and Ostrom – Nobel laureates who made their contributions in an earlier era).

RESP: Thank you for pointing this out, we will take this into consideration for future work.

2. Reviewer 2
   (a) While carefully elaborated and dissected with respect to underlying assumptions and heuristics, it left me with the impression that some of the most challenging aspects (to which the authors allude) relate to the assessment of value alignment. While intuitively challenging, the paper assure the establishment of such function by (admittedly, simply) assuming consensus about about the alignment function. How this comes about is left open. I see two pathways that could be highlighted more explicitly (or, rather, earlier). Instead of discussing the weighing/prioritisation of goals for the purpose of value aggregation in the discussion, why not simply introducing the expectation to weigh values in the first place (as opposed to leaving this to the aggregation functions). This could effectively act as a facilitator to reduce the probability of irreconcilable conflicts later during the alignment assessment.

      RESP: These two suggestions are possible ways to investigate in future work, for now we have left the mechanics of the alignment assessment open. The ultimate discussion lies in the empirical advantages of having to agree only on a few consequences (concrete goals) that are acceptable to all stakeholders, rather than on (abstract) principles. The weighing of values will depend on the value and stakeholder combination - which we have now stated more explicitly in the paper. As mentioned above in a reply to Reviewer 1, consensus depends on the particulars of a given OI but in the extreme case, if there is no effective consensus among stakeholders, the OI wouldn't be viable.

   (b) Secondly (and in addition the first point), to distinguish more explicitly values related to outcome (goal) and process (means), which could, at least conceptually increase the flexibility (and probably operational value) of the aggregation function. This may provide a basis to reduce (while not remove) the need for conflict resolution mechanisms, as an alternative to the consensus assumption (which by design may limit applicability of the concept). To be clear, many of these features are alluded to in the paper (conflicting values, notions of weighing), but could be more clearly elaborated on (or re-positioned in the concept) to signal that the otherwise rigorous assumptions may not be overly constraining in practice as they might read.

      RESP: We thank you for your suggestion and will investigate this in future work. That said, we do not suggest consensus beyond what is said in the previous response, and hope the changes in the text regarding alignment assessment have made this more clear.

3. Reviewer 3
   (a) Minor: WIT (referring to World, Institution, and Technology) model representation, which comes from authors' previous work [IC 2021], is new for me. It would be good to include a citation upfront and refer to it by its full name along with the acronym.

   RESP: We have extended and reorganised the text concerning the WIT and also included clear pointers to previous work.

   (b) Section 4 discusses about values and value alignment problem at length and situating those with online institutions. Section 5 follows up with heuristics to design value-aligned online institutions. The assumptions and the heuristics all read plausible and relevant. These thoughtfully cover various scenarios including value-value conflict, value-goal conflict and goal-goal conflict and how one could prioritise institutions considering the requirements and constraints. After reading these section, I felt a need for a case study that demonstrate how everything ties in together. I think a running example early on and case study would improve readability and understanding of the contributions of the proposal; perhaps the post-proceedings could include that.

   RESP: We have pointed to earlier work for an example, space limits prohibit us from an extensive example in the current paper.