Design heuristics for ethical 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] 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. A major challenge in AI is designing autonomous systems that capture the values of stakeholders, and do so in such away that one can assess the extent to which that system's behaviour is aligned to those values. In this paper we discuss our response to this challenge that is both practical and built on clear principles. Specifically, we propose eleven heuristics to organise the process of making values operational in the design of particular class of AI systems called online institutions. These are governed systems of interacting communities of human and autonomous artificial agents.

Keywords: Online Institutions, WIT Design Pattern, Conscientious Design, Embedding Values, Value Alignment, Value-sensitive Design

1 Introduction

In the Reith Lectures broadcast by the BBC at the end of 2021 [28], Stuart Russell spoke about the challenges Artificial Intelligence (AI) research has in ensuring that AI works for the benefit of human kind. There are several ways to address these challenges. One way is to "put ethics into AI"; and more precisely, focus on the challenge of the **value alignment problem** (VAP): "to build systems whose behaviour is provably aligned with human values". The VAP, in fact consists of two linked problems: how to *embed* human values into AIS and how to assess if, or to what degree, the behaviour of the AIS is *aligned* with those values.

We propose a principled and practical way of approaching the VAP, which we call *conscientious design*, that consists of: (i) *restricting the problem* to one particular type of Artificial Intelligence Systems (AIS) that we call online institutions (OIs); (ii) developing a *conceptual framework* —involving terminological distinctions, formal constructs and properties— that delimit the interpretation of the VAP; (iii) developing *methodological guidelines and heuristics* to guide the embedding of values in an online institution and assessing the OI's alignment with those values; (iv) developing *test cases* which provide both a source of inspiration for the conceptual framework and to evidence how our approach can be put into practice.

This paper is a contribution to component (ii) above. It contains some heuristics that serve to guide the process of making values operational for an OI. The heuristics are intended to be as generic as possible in order to show what are the main practical issues involved in embedding values and assessing alignment. It is work in progress (rather than a completed design methodology) which builds on a decade long research effort investigating online institutions and a *conscientious design* approach for building them successfully (e.g., [34] and see for example references in [1,20,22]). In addition to that long lasting interest, we draw also from experience from a different application of the framework: policy sandboxes, where some of the concepts and constructs involved in the heuristics we present here were first devised [24,25].

Online institutions (OI) are a subclass of artificial intelligent systems. They are *hybrid multiagent systems* (that involve human and artificial participants), where all interactions are *regulated* (only those actions that comply with the OI's regulations can have any effect within that OI), are *online* (interactions consist of messages —or percepts— exchanged through the OI) and, finally, *situated* within a particular socio-technical-legal context [18]. Online institutions capture several intuitions of classical institutions: Searl's notion of separate "crude" and "institutional reality" [30]; North's characterisation of institutions as artificial constraints that articulate agent interactions [23]; Simon's thinking of institutions as an interface between a collective objective and the individual decision-making of participating agents [31]; and Ostrom's criteria for institutional persistence. Those similarities are shown as part of the WIT design pattern in Fig. 1a.

Our focus on online institutions is based on two observations: first, the specific features that distinguish them from other AIS provide the grounds for a principled approach to the VAP; second, plenty of deployed AIS which belong to the OI class and there will be more abundant in the future.

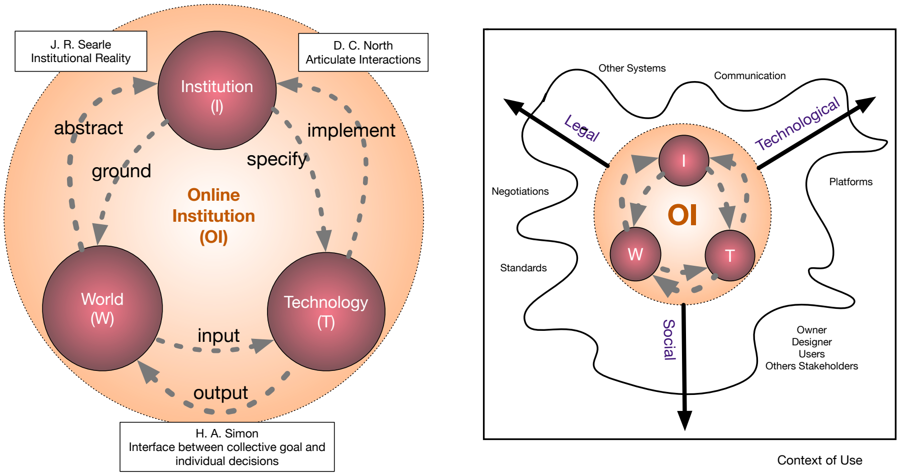
In addition to a precise characterisation of an OI, the main contributions to the conceptual framework of the CD approach are (i) the WIT design pattern, (ii) the (design) distinction between the isolated and the situated view of an OI; (iii) specifying three properties of OIs that one should aspire to achieve in their design: *cohesiveness* (that the three distinctive WIT aspects actually complement each other), *integrity* (that the OI is stable, not corruptible and works as intended) and *compatibility* (with the legal, technological and social constraints of the context where it is situated); and (iv) the proposal of three *conscientious design value categories*: thoroughness, mindfulness and responsibility. Needless to say, appropriate terminological distinctions and some specific constructs give substance to the three main contributions [18].

The main contribution of this paper is to show how each of these four concepts can be translated into methodological guidelines in the form of heuristics for the actual embedding of values. In order to achieve this, the next section provides an overview of our contributions to date. In Sec. 3 we describe a running example to illustrate the applicability of our heuristics. Sec. 4 presents the heuristics themselves and the paper ends with some closing remarks on what we have achieved and future work.

2 The Conscientious Design story so far

2.1 The WIT Design Pattern

The purpose of the $\mathcal{WIT}$ design pattern is to support the process of building online institutions (OIs). The most recent description [22] is a relatively high-level one in-



(a) The Isolated OI, drawing on Searle [30], North [23] and Simon [31]

(b) The Situated OI

Fig. 1: The Views of an Isolated Online Institution vs a Situated Online Institution

tended for a non-specialist audience,while earlier iterations at previous COIN(E) Workshops [20] and other published research [18,19,34] are more technical, and chronicle the evolution of our ideas.

The first significant difference between earlier work, before [22], and the work in this paper is the use of the term *WIT Design Pattern* to refer to the range of concepts and approaches needed for the ethical design of OIs, where we draw on the principles put forward by Alexander [2,3] to capture the idea of habitable *online* spaces that evolve to meet the changing needs and values of their inhabitants. This in turn draws on value-sensitive design (VSD) [8,9,10] which provides the basis for the role of human values in the design process of computational systems, and on Deming's underpinnings for Total Quality Management (TQM) [4] to account for the maintenance and evolution of the online space.

The second significant difference is our use of the term "online institution" (OI) instead of the previously used socio-cognitive technical systems or hybrid online systems. We next describe two distinct categories (or abstractions) of an OI as follows: (i) the *isolated OI* in Fig. 1a, which enables the design of an OI to be considered from three different but related perspectives: $\mathcal{W}$, the OI as seen from the world perspective; $\mathcal{I}$, the institutional or governance perspective of the OI and $\mathcal{T}$, the OI from its technological perspective; and (ii) the *situated OI* in Fig. 1b, where the isolated OI connects with the corresponding elements of the physical and social world to establish what "counts-as" [14] in both directions and to anchor the online institutions with its physical world counterparts.

For any isolated OI it is necessary to be able to demonstrate *cohesiveness*, which is to say the three views work as intended, and *integrity*, which means it is a persistent, well-behaved online system. In order to be fit for its purpose, the situated OI needs to be effective in the context of its use. Consequently, the OI has to be *compatible* with the technological, legal, social and economic requirements of its working environment.

2.2 Conscientious Design Value Categories (CD-VCs)

As part of the development of the WIT-DP framework we have developed the notion of *Conscientious Design value categories*: thoroughness, mindfulness, and responsibility. Here we summarise these to provide the reader with a sense of these below (the full definitions can be found in [22]):

- **Thoroughness**: this refers to conventional technological values that promote the technical quality of the system. It includes completeness and correctness of the specification and implementation, reliability and efficiency of the deployed system. Concepts such as robustness, resilience, accessibility, and security are all aspects of thoroughness.
- **Mindfulness**: is about engendering a wider awareness of the range of direct and indirect needs of, and impacts on, humans (both users and non-users) which is so often over-looked. Examples include data ownership, and the OIs accessibility and usability, and this category has much in common with Schwartz' "personal focus" values.
- **Responsibility**: addresses both the effects of the system on stakeholders and the context in which it is situated, as well as how indirect stakeholders and that context may affect internal stakeholders. Examples include liability and prestige, and are akin to the "social focus" values of Schwartz [29].

In our work on Ethically Aligned Design [22] we have shown how these CD value categories can be mapped onto different ethical AI value frameworks such as the initiatives from the EU [11] on Trustworthy AI and the IEEE guidelines for imbuing values in AIS [32]. As meta-analyses of the multitude of frameworks show [7,17], many have overlapping definitions and principles. However, the CD value categories have the advantage of supporting more than one way of looking at the principles included in these frameworks.

One final remark here concerns the stakeholders. Stakeholders are all those affected by, or those affecting, the system during both development and deployment. *Direct* stakeholders are those stakeholders who are responsible for the design and deployment, or are direct users of the OI. In practice, in every OI there are always three categories of direct stakeholders: owner, engineer and user and we will detail each of these in the next section. Those stakeholders who are affected by the system, but are not part of the decision-making and do not use the system directly, we call *indirect* stakeholders — as is the usual term in value sensitive design. The values of direct stakeholders need to be explicitly accounted for in the design and use of the OI.

In order to identify those values of direct stakeholders and make them operational, direct stakeholders can be separated in three different groups: *owner, engineer and user*. This separation reflects the distinctive objectives of direct stakeholders in every OI: the owner looks to deploy an OI that supports a collective endeavour "as well as possible", the users participates in the OI to achieve "as well as possible" their individual goals with whatever means are provided by the OI, and the engineer builds “as well as possible” an OI that satisfies “as well as possible” the owner and the user objectives. The point is that each “as well as possible” is guided by different values. Notice that since, in every OI, those distinctive objectives of each of the direct stakeholders are similar, the values that each of them holds are similar to some extent in every OI. See below, Sec. 4.2, Heuristic 4.

3 The Easyrider Online Institution

To support the understanding of the theoretical and practical concepts involved in the WIT-DP for ethical AIS, we introduce Easyrider, a rich enough toy example of an OI for buying and selling train tickets online. Are we mentioned in the last section the three categories of direct stakeholders are Owner, Engineer and Users and are detailed as follows.

1. Owner: refers to the individual or organisation that commissions and operates the OI. In this case the railway company is the Owner, because it commissions and operates the OI in order to sell tickets online through travel agencies.
2. Engineer: refers to the individual or organisation responsible for ensuring the requirements of the owner are satisfied in am effectively designed and deployed OI that supports intended usage.
3. User(s): refers to the users who will use the system and satisfy their goals by interacting with others. In Easyrider there are two categories of users: *passengers* (who are human agents) that use Easyrider to buy, and possibly return train tickets, and *travel agencies* (who are software agents) that buy tickets from the railway company to re-sell them to passengers.

In Easyrider, the indirect stakeholders would include the commerce and transit authorities that regulate the railway services, the banks and payment services that support purchases, phone companies and, to some extent, the population —and the environment— of those cities served by trains and affected by the travelling of people back and forth.

3.1 Goals and Values

The WIT DP approach to design we propose starts by identifying the ultimate objectives of stakeholders —the rationales for the creation, engineering, and use the particular OI. However, because we want to embed values in the OI we also need to make explicit the *terminal* (or intrinsic) values that motivate those objectives and those *instrumental* values that determine the means provided by the OI to reach those objectives [27].

Table 2 illustrates those three elements in Easyrider. For brevity, we only include the *ultimate goal* of the stakeholder groups, the key terminal values that guide those goals and the most prominent instrumental values that motivate the stakeholders’ decisions and means to achieve those goals. Next to each “instrumental value” we indicate the type of CD category it belongs to (T for thoroughness, M for mindfulness, and R for responsibility). In the next section we build on these examples to illustrate how CD values can be embedded in Easyrider.

For example, the railway company who owns Easyrider develops an online ticketing service in order to sell enough seats to amortize capital it has invested in the train service, and it wants to achieve that objective guided by three terminal values: (i) a sense of good management of the company capital and its operation; (ii) the provision of a service through travel agencies that is profitable for these travel agencies which in turn leads to attracting both existing and new passengers to use the system; and (iii) an acknowledged positive impact because more persons travel in train instead of using less ecological means of transportation and also because a public infrastructure is better used.

Moreover, the specification of Easyrider should also reflect the railway company's criteria for instrumenting those terminal values. So, for instance, good management is achieved by a thorough implementation of management policies and practices; responsibly by achieving a healthy cash-flow. Alongside, the OI promotes an occupancy of wagons that provides that cash-flow without being uncomfortable for passengers; while enabling profitable margins to travel agencies.

We now move onto the issue of how to make values operational within our established framework for designing ethical OIs.

4 Making values operational

The proof of developing a value-imbued system is in the pudding of making values operational as well as choosing the values in order to be able to assess if the values are indeed enhanced or supported by the system. According to [26], there are three pre-requisites that need to be fulfilled to assess if certain values are embodied in an AI system: (i) values are addressed in the design of the system, i.e., there is no such thing as accidental value embedding; (ii) the AI system is seen as a sociotechnical system not an isolated technological artefact, i.e. it is situated; and (iii) the AI system is not ascribed any moral agency, differentiating it from human agents.

Since we want to embed values in a working system, we need to translate an intuitive understanding of values into precise constructs that can be specified as part of a system and then see whether or not they are supported by the working system. This is what we call the process of making values operational. Since this is a complex process the first thing to do is to make things manageable.

4.1 Three Heuristics for structuring value operationalisation

The point of the heuristics for structuring value operationalisation is threefold: (i) to decompose the complex problem into subtasks, (ii) to facilitate the separation of design concerns and (iii) to put design priorities in focus. We propose three design heuristics for this purpose:

Heuristic 1. **Making values operational** is an iterative process.

Making values operational is a process of iterative approximation that converges to whatever is "just enough" for whichever stage the system has reached, from preliminary evaluation through to decomissioning. It also functions as the means to track the moving

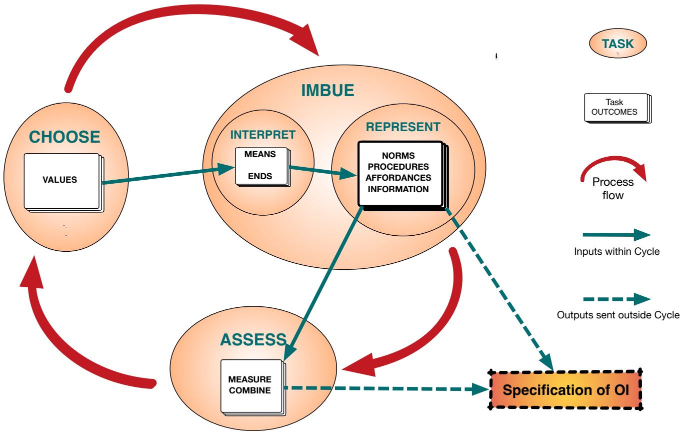


Fig. 2: The process for bringing values into the design of an OI

target of the changing needs and value preferences of the participants. As sketched in Fig. 2, the process starts with the choice of values and ends with a specification of an OI that is aligned with those values. The first task consists of choosing a *list of values* that are relevant for the OI. The task of the second stage is to make those values objectively measurable, for which we use a two step process: they are *interpreted* by linking them to concrete referents (“means” to support the value and “ends” that reflect its achievement) that may then be *represented* within the system in readiness for the next stage. The third stage consists in defining the *value assessment models* that establish (i) the precise ways in which one can tell whether a value is being attained and to what degree, and (ii) how to resolve value conflicts . The outcome of this process is to put the representation of the values and the assessment into the specification of the OI.

Heuristic 2. **Ethical design is a participatory effort** where all direct stakeholders have their say at different phases of the OI life-cycle.

The cycle of making values operational is active for the lifetime of the OI. However, the involvement of stakeholders is different in different phases of that life-cycle. The design of a value imbued process is started by the owner whose main goals and values are passed as design requirements to the engineer. The engineer is then responsible for interpreting these values of the owner, and to elicit and interpret the values of users. Based on these requirements, the engineer makes all the relevant values operational and specifies and deploys the system as proficiently as possible. Although the decision to deploy rests with the owner and their values take priority, its success rests with the users and in the implementation. Therefore, in the evaluation and updating of the system, user values take precedence, then the engineer takes over and the release of a new version is up to the owner’s values again.

In practice (as mentioned in Sec. 3.1), the process of making values operational is kick-started by the choice of *terminal values* (desirable end-states of existence) for the ultimate goals of each stakeholder and a first take on the *instrumental values* (related to modes of behaviour) [27]. In other words:

Heuristic 3. **Value assessment** drives the iterative process of making values operational.

The rationale is that it is helpful to sketch which are the values that each stakeholder wants reflected in the OI and how stakeholder would assess whether the OI promotes or protects those values before starting the detailed process of imbuing values.

4.2 Heuristics for the choice of values

A first heuristic is based on the acknowledgement that the choice of values needs to take into account three frames of reference. First, the *application domain*, which determines goals and makes some instrumental values relevant and others less so. In Easyrider for example, values related to e-commerce and transportation become relevant, while those associated with, say, health services do not. Second, the *role of stakeholders* influences the choice of values. Stakeholders choose values that are relevant for the domain, however, regardless of the application domain, engineer values always reflect the goal of developing an OI that handles a particular collective activity , owner values always have to reflect the need of engaging users, and user values reflect their motivation and preference for choosing to engage in the OI. The third frame of reference that influences the choice of values is to profit form the fact that the *WIT design pattern* induces a natural separation of design concerns that remain valid throughout the OI life-cycle.

Regarding the use of the WIT design pattern, we argue that in order to embed the terminal and instrumental values of each stakeholder in the OI, one needs to address three main design requirements: (i) to enable collective interaction in a well-defined, limited part of cyber-physical reality; (ii) to set up the rules of the game so that the outcomes of those interactions are consistent with the values of the stakeholders; and (iii) to implement these rules in such a way that the actual online system runs according to those rules. The $\mathcal{WIT}$ pattern facilitates the analysis of those requirements by establishing **nine design contexts** where specific values are involved. These contexts are the six design concerns associated to the relationships between the $\mathcal{W} - \mathcal{I} - \mathcal{T}$ components of the isolated OI (Fig. 1a) and the three design concerns arising from the legal, technological, and social compatibility of the situated OI (Fig. 1b). Two points are worth mentioning: first, all CD, terminal, and instrumental value labels may be localised as more specific labels for each stakeholder in each of the nine contexts; second, not all the nine contexts are equally important for all stakeholders, hence one can rank the degree of involvement —in the participatory design process— of the three stakeholders for each context and each CD value class.

Table 1 illustrates value contextualisation for the OI engineer regardless of the OI domain. The top part gives an interpretation of the CD-value categories and the bottom part declares those contexts of the WIT design pattern where the engineer has the final word on the choice and interpretation of the contextualisation.

Heuristic 4. **Contextualisation:** Value choice depends on the domain of the OI, the actual stakeholder and the WIT-DP context where it is meant to be applied.

The second heuristic for choosing values suggests how to proceed in order to identify relevant values. The idea is quite straightforward: use the goals of the stakeholders to search for values and keep the CD value categories present to prevent overlooking a significant value.

Heuristic 5. **Value selection:** Define the *ultimate goals* of each direct stakeholder, then associate with each stakeholder the corresponding *terminal* and *instrumental* values and validate the selection of instrumental values with the *CD value-categories*.

In practice, each stakeholder is committed to an ultimate goal which ought to be legitimised by an ultimate or intrinsic value. However, that goal needs to be decomposed into means and ends that determine how the stakeholder may achieve its goal. In order to choose the particular means and ends that lead to achieving that ultimate goal the stakeholder will use its instrumental values [5]

The CD value categories serve a dual purpose, on one hand they are useful for labelling instrumental values (something that will be essential for value assessment and for the eventual termination of the operationalisation process); on the other hand, the intuitive understanding of the three categories (and the experience of using them in other OIs) is a practical way of validating that the instrumental values that have been chosen truly constitute a good coverage of each of the three main categories.

Table 2 is a partial contextualisation of the terminal and instrumental values of the owner and the users of Easyrider (the engineer's values are summarised in Table 1). In Table 2 we list only four instrumental values of the owner and users of Easyrider, and refine these with more specific values; some of which are underlined because they are used in sections 4.3 and 4.4, and in Table 3 to illustrate the interpretation and representation of value labels. Notice that each instrumental value is labelled with the CD value category it more naturally belongs to.[6]

One last remark about the choice of values. Since the process of making values operational is gradual, the refinement of value labels is better served by the analysis of only the most salient stakeholder values in the first pass. One need only come back to this step of the operationalisation process when the value assessment process requires an improvement of the alignment of the OI to the stakeholders' values (see Heur. 11).

4.3 Heuristics for value imbuing

Imbuing is a prerequisite for specification. Its objectives are to turn the *intuitive* understanding of a relevant value into an *objective* understanding that may be embedded into the OI. This task of imbuing values in a system involves two efforts: *interpretation* and *representation* of values. These two sub-processes are applied to each instrumental value label and while all stakeholders are involved, the stakeholder who chooses a given value leads the task.

[5] There are two ways of identifying ultimate and instrumental value labels. One is to ask the users to name them [33,15]; another is to draw from available value taxonomies like [12,13,29]. Following the second path, we propose the CD value categories mentioned earlier: *thoroughness*, *mindfulness* and *responsibility* [20,22] that serve as intuitive catch-all labels that become more meaningful as they are applied to different design concerns as the design of the OI advances.

[6] Although individual passengers and travel agencies may have different value interpretations, the table stands for a consensus of what values to embed and how that is the result of the participatory design process.

(1) Engineer’s terminal and instrumental values

Engineer’s ultimate goal: Design and build an OI proficiently

Thoroughness:
(i) Do the usual stuff to do a good job during the whole life-cycle of the system;
(ii) Adopt best practices and standards in the application domain;
(iii) Make the OI fit for the ultimate goals of direct stakeholders;
(iv) Validate cohesiveness and integrity.

Mindfulness:
(i) Engineer all values of owner and users;
(ii) Be transparent about the quality and limits of the OI.

Responsibility:
(i) Guarantee cohesiveness and integrity of the isolated OI;
(ii) Guarantee compatibility of the situated OI.

(2) Engineer’s leadership in the WIT design pattern:

(i) Integrity of isolated OI;
(ii) Cohesiveness of isolated OI;
(iii) Technological compatibility of situated OI;
(iv) Priority design sub-contexts: specification (I→T), implementation (T→I) and user interface (W↔T’)

Table 1: Engineer’s value contextualisation (independent of OI domain). (1) The generic ultimate goal of an engineer is aligned with each of the CD-value categories, which are translated into intuitive descriptions of their most salient means and ends. (2) The engineer holds the ultimate responsibility for value imbuing in particular WIT pattern design contexts.

Railway company	Passengers	Travel agencies
Fill trains	*Buy train tickets*	*Profitable trading business*
Sound management adequate return on investment (M), balanced cash-flow (M), ...	**Convenience** flexibility (M), abundant offer (M), ease of use (M)	**Profit** increase volume (M), increase margin (M), lower costs (M,R), ...
Proficient OI trustworthiness, (R) effectiveness (M, R), impartiality (R), transparency (R), legal compliance (M,R), ...	**Restraint** lower fares (M), ...	**Convenience** easy to use (M), compatible with in-house practices and systems (M), reliable support (M), ...
Good customer relations reliable support (R), accountability (R), privacy (R), ...	**Reliability** secure transactions (M), accountability (M), privacy (M), ...	**Reliability** transparent rules of operation (M), fair competition (M, R), secure transactions (M,R), ...
Good citizenship support SDGs (R), corporate responsibility (R), prestige (M), ...	**Pleasant travelling** comfort (M), conviviality (M,T), ...	**Good citizenship** prestige (M), social recognition (R), ...

Table 2: Ultimate goals and main instrumental values of the owner and users of the Easyrider OI. Each goal is associated with four instrumental values that guide its achievement. Those instrumental values are in turn partially refined into more specific values – labelled with their corresponding CD-categories – that will be imbued in the system. Underlined values are used in Table 3 and examples.

1. Interpretation: Its purposes are to obtain an objective description of the the mechanisms and constraints that support (promote) or maintain (protect) each value, and an objective description of how one can eventually assess whether a value is in fact being protected or promoted. This can be articulated with two heuristics.

Heuristic 6. **Value interpretation (1)** is to articulate the meaning of a value as the *means* and *ends* that are most typical of it in a given context.

The leading stakeholder for a given value, with inputs from the other stakeholders, interprets it by looking at the concrete actions or objects that can afford its achievement and maintenance (the *means*) and identifying the states of affairs that show that the value is actually being promoted or protected (*ends*).

Once the means and ends are articulated, one needs to identify what the observable features of the states of affairs are involved in those means and ends in order to use them for measuring the attainment of a value and stating along those terms the degree of satisfaction of the different stakeholders. Consequently, this heuristic provides the essential elements for the definition of the value assessment models that we discuss in the next section.

Heuristic 7. **Value interpretation (2)** consists in associating with each value *observable features* involved in value means and ends, and discovering stakeholder priorities and thresholds of satisfaction.

2. Representation: From these means and ends, and their observable features, the engineer with input from the other stakeholders decides how to *represent* the instrumental values so that they can be implemented as part of the physical and governance model of the OI (or in the decision model of an artificial agent).

Heuristic 8. **Value representation** translates value interpretations into components of the abstract representation of the OI, that will be the basis for its specification.

There are essentially three ways of translating value interpretations into value representations: as norms and standard procedures, as affordances, and as information for participants. Table 3 illustrates the interpretation and representation of some instrumental values included in Table 2).

1. Some values are represented directly as *norms* that promote, mandate, curtail, or discourage behaviour; or prescribe the consequences of institutional actions. For example, passengers' flexibility may be *interpreted* as allowing ticket changes, which may be *represented* with a norm that allows ticket purchase and devolution up to five minutes before departure.
   Sometimes a single norm is not enough and a value may have to be represented as a standard procedure. For instance, Easyrider may include protocols for issuing different reports. Such reports —say, tax-valid receipts for every final sale or a refusal to accept a devolution—, are *means* that support the *end* of having evidence to achieve the value of accountability and transparency for stakeholders.
2. A second way of going from interpretation to representation is through the introduction of new entities in the institutional reality that *afford* specific actions or

	Passenger	Users and owner	Owner	Owner
Values	flexibility	accountability and transparency	support SDGs	adequate return on investment
Ends	allow last-minute purchases, ...	proof of action, ...	promote the use of train to support SDG 7, 9, 13, ...	high occupancy of carriages
Means	extend purchasing deadlines; install ticketing machines at station; ...	reports of relevant transactions, ...	marketing campaigns, ...	attractive fares, ease of purchasing, marketing, ...
Representation	norms and affordances	procedures for issuing each report type	banners and messages, poll	procedure and physical action; add carriages when needed
Observable	number of tickets sold close to departure; number of machine-issued tickets	list of reports of each type	passenger and TA awareness of the good impact of trains	occupancy level
Thresholds	more than 10% of total sales are late purchases	at least all legally required reports	increase of awareness and acknowledged motivation	between 60% and 80% occupied seats in a carriage

Table 3: Imbuing of some instrumental values of Easyrider's owner and users. Each value is interpreted in these examples as one typical end that leads to the stakeholders' ultimate goals in alignment with the corresponding values, and some means that are conducive to the achievement of that end. These means and ends would be represented with some instruments that embody the means, in a way that one may objectively assess whether these values are satisfied or not in the deployed system.

outcomes that promote or protect a value like accountability. For example, passengers' value of travel flexibility may also be supported by allowing the possibility of purchasing and printing tickets in ticket dispensers at the station. In this case the physical model (of $\mathcal{W}$) would need to include ticket dispensers and their use would be regulated with norms that will be part of the "governance model" of Easyrider. In this example, the *affordance of using printed tickets* may require other devices in the station or aboard trains to validate tickets. The owner would have to decide whether the use of printed tickets is worth the extra regulations and the cost of dispensers, or not.

3. The third mode of representing values is as a set of facts, recommendations or arguments that are made available to users with the purpose of influencing their decision-making. For example, the railway company's instrumental value support sustainable development goal (SDG) can be promoted through banners or messages that appear in the use of Easyrider or in marketing campaigns that make users aware of the beneficial impact of traveling by train (and eventually also increase the number of trips). The achievement of the value is observable, for example, through a customer satisfaction poll and its degree of satisfaction measured through the aggregate opinion users.

4.4 Heuristics for value assessment

We now turn our attention to the task of evaluating to what extent stakeholders values are reflected and met in the OI. The imbuing step that we proposed above entails three claims: (i) that —since ends are observable— the alignment of values can be "assessed" somehow (or *measured*); (ii) that stakeholders are capable of determining whether they are satisfied or not with the degree to which the system is aligned with the values they care about —since for each value interpretation, its satisfaction thresholds can be elicited from stakeholders; (iii) that the engineer is able to transcribe measuring and satisfaction into the specification of the OI. We make these claims operational with the construct of *value assessment models*. The value assessment model of a stakeholder *s*has three parts: a list of values, a way to measure each of those values, and a way to combine them.

Heuristic 9. **Value measurement** consists of mapping the observable outcomes that stand for the value and the thresholds expressed by the stakeholder on an ordered set that reflects the degree of satisfaction of the user with that value.

We mention two extreme possibilities of value measuring to illustrate this heuristic. As we saw in the previous section, the interpretation of a value commits to an observable feature that stands for the value and, ideally, to some bounds or thresholds that determine the degree to which the value is satisfied. one form of measuring values that allows for a crisp assessment assumes that the observable feature is an "indicator" (or a scale on a totally ordered set), boundaries determine thresholds that determine not only if the value in question is being satisfied or not but also to what degree.[7] For instance, in Easyrider, a travel agency recognises secure transactions as a mindfulness and responsibility value, which is being interpreted as "honouring deals". This instrumental value is interpreted, in particular, by guaranteeing that travel agencies pay all their dues to the railway company and to other travel agencies. The means the institution has implemented to maintain that value, are to require of travel agencies to post a bond that covers potential harm, and levy a fine for any mishap. The observable outcomes are the costs of the mishaps. The travel agency may use that representation of the value to measure secure transactions and also the satisfaction of its own value of lower costs by the sum of fines it pays over the year and prefer to pay no more than a fixed amount in a year.

A minimalistic way of measuring value satisfaction, on the other hand, may consists simply in mapping all the possible observable outcomes onto a finite set of proxy scores that are each labelled with a degree of satisfaction that reflect the boundaries defined in the interpretation of the value. For example, in Easyrider, the railway company wants to fill trains but not too much if it wants to keep passengers satisfied. The owner satisfaction depends not only on the number of unsold seats (few sold seats, not good; totally full trains, not good either), but also in how the empty seats are distributed in each carriage (few passengers but all stuck at the back, not good; groups of friends

[7] Ideally, the totally ordered set is mapped onto a convex function whose range goes from -1 (totally unsatisfied) to 1 (perfectly satisfied) and the mapping of thresholds define a region of "satisficing" scores.

seated together, good). Satisfaction of passengers' comfort and conviviality as well as affecting the railway company's balanced cash-flow could be measured, for example with a pairwise preference combination of density vs seat configurations and the degree of satisfaction of each pair with a ranking, say, unacceptable, undesirable, satisfactory, very satisfactory. Even more radical, the value accountability may be interpreted as responsibility by the owner and in this case, if the same bonding mechanism is afforded, its fulfillment duly regulated and its enforcement strict – all these conditions achievable at implementation time – its assessment is *ex-ante* satisfactory.

The third component of the value assessment model is an aggregation function that combines the stakeholder's satisfaction with all and every value; and thus assess the extent to which the OI aligns with the combined set of stakeholder's values. The aggregation function should take into account the priorities and trade-offs between values and other features like their urgency, associated costs or expected evolution of the observable features involved with those values.

Heuristic 10. **An aggregation function** combines the level of satisfaction of several values into a single outcome that represents the aggregate satisfaction derived by the stakeholder from the combination of those values.[8]

A thorough discussion of aggregation functions is beyond the scope of this paper but one can get an idea with a simple version of the engineer's aggregation function. A top-down definition of the engineer's aggregation function may be to aggregate the degree of satisfaction of the engineer with each of its three CD values defined in Table 1, as follows: (i) Assessment of satisfaction of thoroughness and responsibility is essentially technical. The first will be the result of the aggregation of the degrees of satisfaction of the four thoroughness goals and by assessing that mindfulness, responsibility, integrity, cohesiveness and compatibility are dully validated. (ii) Likewise responsibility is assessed through the assessment of the (technical) soundness of integrity and compatibility of the OI. (iii) However, satisfaction of mindfulness requires that all the values of users and owner have been properly "engineered" (specified and implemented) but for that owner and users have to agree on the way their values are interpreted and represented. Thus engineer's mindfulness depends on users and owner agreeing that their own values of throroughness, mindfulness, and responsibility are satisfied with the observable features and thresholds that they agree upon.

This very last aggregation involving the satisfaction of the other stakeholders builds on the process of participatory design of the OI and on the assessment of each separate value in terms of the observable feature that stands for it (which is the same for every stakeholder). The way these detailed assessment are aggregated may be different for each stakeholder but in this case, the engineer has priority on some CD design contexts (part b in Table 1) and thus its aggregation function of non-priority context will be that of the other stakeholders but the engineer's may be more demanding for the values in its own priority contexts. The owner, as the stakeholder who is responsible for

[8] Note that to determine the alignment of an OI with a set of values, which is the ultimate purpose of making values operational, one needs a top level aggregation function that combines the degrees of satisfaction of all stakeholders.

commissioning, deploying, updating and preserving the operation of the OI, has the last word.

Note that the purpose of the aggregation function is two-fold: first to commit to an encompassing measure of satisfaction that reflects value priorities and trade-offs for the stakeholder; second to determine if the alignment of the OI with the set of values is "good enough" for the stakeholder. Consequently, if the alignment is not good enough, the aggregation function and the value assessment model in general can be used to pinpoint those values that are not properly embedded in the OI. If a global assessment model is not satisfactory, a compromise can usually be reached by revising the aggregation function, simplifying value measurement, and relaxing satisfaction thresholds.

Heuristic 11. **Improvement of value alignment**. When a value alignment is not satisfactory, revise the steps of the operationalisation process backwards until stakeholders are satisfied.

The idea behind this heuristic is the following: from a bottom up perspective, each stakeholder chooses its own values, how to interpret them, and the observable features that are used to determine whether the value is being satisfied (and to what degree) (Heuristic 7). One underlying assumption of OIs is that there are observable features which are common to all stakeholders. However, not all stakeholders will hold the same values in general, and therefore not all observable features will be equally relevant for different stakeholders. This means that each stakeholder will combine and prioritize the observable features in different ways. This difference, is unproblematic unless a conflict of the interpretation and assessment of values among stakeholders arises. When this occurs, the conflict can be resolved by incorporating additional observable features (and the new required means to achieve them) that are relevant for the stakeholder who is unsatisfied with a specific interpretation of a value into means, ends and observable features.

From a top-down perspective, we can assume all stakeholders aggregate values in our three CD categories: thoroughness, mindfulness and responsibility. The aggregation function of each stakeholder is unlikely to be the same in general, and agreement, or some other form of reconciliation should take place, in order to the the OI to be aligned with each of its stakeholders values. This is unproblematic as long as the stakeholders agree on some trade-offs which may be reached if some stakeholders change the weighting of some values in the aggregation function, or choose to relax their levels of satisfaction with respect to certain values.

The final trade-off agreement may be reached by moving back and forth from the aggregation at different levels of value decomposition within each category.

5 Closing remarks

In this paper we propose heuristics to make stakeholder values operational in online institutions. These heuristics belong to a larger task to provide general methodological guidelines for a principled approach to embedding values in AI systems. It seems clear to us that any such approach requires that values are made explicit, that their interpretation can be translated into a machine executable representation, and that their satisfaction can be objectively assessed. We claim that while these conditions are necessary, we do not impose any further requirements to value theory.

In the heuristics we propose, we remain neutral about the choice of formalisms used for representation and for the assessment of values. (Though we are considering using Z with its ability to capture both agent architectures, multi-agent systems and design methodologies [5,6,16].) However, we believe that for certain types of online institutions (and AIS in general) there are reasons to adopt specific interpretations of each value in terms of a means and ends decomposition that give grounds to more specific representation and assessment conventions, whilst recognising they might not necessarily be unique.

Whilst focus of this paper has been on heuristics for making values operational in governed multi-agent systems, we believe that heuristics could be similarly applied to the embedding and assessment of values in the design of individual autonomous agents. Nevertheless, there are specific aspects of the design process that would need to address the role of values in designing artificial agents' architectures and behaviour. For instance, for an artificial agent that is intended to behave in an ethically-consistent manner, the engineer may commit to some cognitive architecture that includes values as an explicit and necessary construct in their inference-based decision-making models, or make explicit use of value theories that explain ethical behaviour without assuming rational ethical reasoners [24].

We mention elsewhere [21] that one could apply the conscientious design approach to developing tools to prevent undesirable effects of existing third party software. The heuristics we propose in this paper can be used to determine whether the behaviours of a given system is aligned with any explicitly stated values. This leads us to the possibility of adding, to such existing systems, new functionality that ensure they behave with proper alignment with respect to any stated values. This is something we plan to address in future work. In addition, our intentions include developing our approach to support policy makers, evolving stronger good practices, and making use-cases readily available to facilitate uptake.

The process of making values operational that we discuss in this paper is at the core of the Value Alignment Problem, which concerns the embedding of values in artificial autonomous systems and assessing their alignment. However, our proposal can be placed in a wider perspective of developing a theory of value with a distinctive AI flavour. The value theory we foresee would be centered on the *interplay of governance, autonomy, and collective hybrid behaviour* and because artificial autonomous entities are involved, there are meta-ethical, normative ethics, and applied ethical problems that other theories of values do not address. In fact, unlike other theories of value, such an "artificial axiology" purports to embed ethical constructs into artefacts and assess ethical questions associated with them. The approach we envision shares with AI and other sciences of the artificial a peculiar mix of science and engineering; it would draw on constructs and methods from AI and other sciences of the artificial, and require a serious interdisciplinary effort.

Acknowledgements. D'Inverno thanks IIIA-CSIC Barcelona for hosting him during a research sabbatical and his host institution Goldsmiths, University of London for making that possible. Noriega's work is supported in part by the CIMBVAL project (Spanish government project #TIN2017-89758-R).

References

1. Aldewereld, H., Padget, J., Vasconcelos, W., Vázquez-Salceda, J., Sergeant, P., Staikopoulos, A.: Adaptable, Organization-Aware, Service-Oriented Computing. IEEE Intelligent Systems **25**(4), 26–35 (7 2010). https://doi.org/http://doi.ieeecomputersociety.org/10.1109/MIS.2010.93
2. Alexander, C.: A pattern language: towns, buildings, construction. OUP (1977)
3. Alexander, C.: The timeless way of building, vol. 1. New York: OUP (1979)
4. Deming, W.E.: Quality, productivity, and competitive position. MIT Press (1982), but see also https://en.wikipedia.org/wiki/Total_quality_management, https://en.wikipedia.org/wiki/Kaizen, and https://en.wikipedia.org/wiki/Eight_dimensions_of_quality
5. d'Inverno, M., Luck, M.: Development and application of a formal agent framework. In: First IEEE International Conference on Formal Engineering Methods. pp. 222–231 (1997). https://doi.org/10.1109/ICFEM.1997.630429
6. d'Inverno, M., Luck, M., Noriega, P., Rodriguez-Aguilar, J.A., Sierra, C.: Communicating open systems. Artificial Intelligence **186**, 38–94 (2012). https://doi.org/https://doi.org/10.1016/j.artint.2012.03.004
7. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M.: Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Tech. Rep. 2020-1, Berkman Klein Center Research Publication (2020)
8. Friedman, B.: Value-sensitive design. Interactions **3**(6), 16–23 (1996)
9. Friedman, B.: The ethics of system design. Computers, Ethics and Society pp. 55–63 (2003)
10. Friedman, B., Hendry, D.G., Borning, A.: A survey of value sensitive design methods. Foundations and Trends in Human-Computer Interaction **11**(2), 63–125 (2017)
11. High-Level Expert Group on Artificial Intelligence (AI HLEG): Ethics Guidelines for Trustworthy AI (2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
12. Hofstede, G., Hofstede, G.J., Minkov, M.: Cultures and Organizations - Software of the Mind: Intercultural Cooperation and its Importance for Survival (3. ed.). McGraw-Hill (2010)
13. Inglehart, R.: Human beliefs and values: A cross-cultural sourcebook based on the 1999-2002 values surveys. Siglo XXI (2004)
14. Jones, A.J.I., Sergot, M.: A Formal Characterisation of Institutionalised Power. Logic Journal of the IGPL **4**(3), 427–443 (06 1996)
15. Liscio, E., van der Meer, M., Siebert, L.C., Jonker, C.M., Mouter, N., Murukannaiah, P.K.: Axies: Identifying and evaluating context-specific values. In: Proceedings of the 20th international conference on autonomous agents and MultiAgent systems. pp. 799–808. International Foundation for Autonomous Agents and Multiagent Systems (2021)
16. Luck, M., d'Inverno, M.: Structuring a z specification to provide a formal framework for autonomous agent systems. In: Bowen, J.P., Hinchey, M.G. (eds.) ZUM '95: The Z Formal Specification Notation. pp. 46–62. Springer Berlin Heidelberg, Berlin, Heidelberg (1995)
17. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. In: Ethics, Governance, and Policies in Artificial Intelligence, pp. 153–183. Springer (2021)
18. Noriega, P., Padget, J., Verhagen, H.: Anchoring online institutions. In: Casanovas, P., Moreso, J.J. (eds.) Anchoring Institutions. Democracy and Regulations in a Global and Semi-automated World. Springer (2022), in press.

19. 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
20. Noriega, P., Sabater-Mir, J., Verhagen, H., Padget, J., d'Inverno, M.: Identifying affordances for modelling second-order emergent phenomena with the *WIT* framework. In: Autonomous Agents and Multiagent Systems - AAMAS 2017 Workshops, Visionary Papers, São Paulo, Brazil, May 8-12, 2017, Revised Selected Papers. pp. 208–227 (2017)
21. 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)
22. Noriega, P., Verhagen, H., Padget, J., d'Inverno, M.: Ethical online AI systems through conscientious design. IEEE Internet Computing **25**(6), 58–64 (2021)
23. North, D.: Institutions, Institutional Change and Economic Performance. CUP (1991)
24. Perello-Moragues, A., Noriega, P.: Using agent-based simulation to understand the role of values in policy-making. In: Advances in Social Simulation. pp. 355–369. Springer (2020)
25. Perello-Moragues, A., Noriega, P., Popartan, A., Poch, M.: On three ethical aspects involved in using agent-based social simulation for policy-making. In: Ahrweiler, P., Neumann, M. (eds.) Advances in Social Simulation. pp. 415–427. Springer, Cham (2021)
26. van de Poel, I.: Embedding values in artificial intelligence (AI) systems. Minds and Machines **30**(3), 385–409 (2020)
27. Rokeach, M.: The nature of human values. Free press (1973)
28. Russell, S.: Living with artificial intelligence (Dec 2021), https://www.bbc.co.uk/programmes/b00729d9/episodes/downloads
29. Schwartz, S.H.: An overview of the Schwartz theory of basic values. Online readings in Psychology and Culture **2**(1), 11 (2012)
30. Searle, J.R.: The Construction of Social Reality. Allen Lane, The Penguin Press (1995)
31. Simon, H.A.: Models of man; social and rational. Wiley (1957)
32. 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
33. Umbrello, S., Van de Poel, I.: Mapping value sensitive design onto ai for social good principles. AI and Ethics **1**(3), 283–296 (2021)
34. 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)