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. The importance of ensuring that the different values of stakeholders are clearly manifest in the behaviour of software systems has been recognized for some time. In the case of AI, ethically aligned reasoning - that often includes consideration of values - is increasingly seen as a way to address concerns at societal and governmental level about its potential negative impact. One major challenge in AI research, is designing processes to explicitly capture the values of stakeholders so that they can be appropriately considered both throughout any design and deployment process. Without an approach to "making values operational", which includes being able to assess the extent to which a system's behaviour is aligned to those values, it is difficult to see how any system can be said to truly embody the values of its stakeholders. In this paper we develop a methodology to address this challenge, continuing from previous work on Conscientious Design and the WIT (World-Institution-Technology) model. We present a set of heuristics to make explicit the relevant human values of stakeholders, embed those values in the operation of online institutions, and to continually assess the extent to which the operation of an online institution is consistent with those values.

Keywords: Online Institutions, WIT Design Pattern, Conscientious Design, Embedding Values

1 Introduction

In the Reith Lectures broadcast by the BBC at the end of 2021 [24], Stuart Russell spoke about the challenges AI research has in ensuring it works for the benefit of human kind. He was quite clear that this is a problem that we take lightly at our peril and was arguably the single most pressing problem that humanity faces in current times, Covid-19 notwithstanding.

There are several approaches to these challenges. One is to design AI systems according to certain values, principles or ethics and then put in place mechanisms and evaluations to evidence that the AI system is clearly working for the benefit of humankind (or some specific subset). A good example of this approach working in practice is the AI4people framework and its core principles [6]. Another approach to this challenge is not focusing on the AI system itself, but on the design of online systems which bring AI and humans agents together, and ensuring that the way the interactions take place is aligned with certain values or ethical principles.

It is this second approach that this paper is concerned with. Namely, how we design and deploy a *governed* online system to facilitate the interaction of hybrid (AI and human agents) communities according to a set of values. For several years we have been researching a specific class of such systems called online institutions [15,17]. The challenge, from an ethical perspective, is to be able evidence to what extent the online system's governance mechanisms ensure that the behaviour of its interacting can be said to be consistent with any given set of values. It is this challenge that we address in this paper.

Over the last couple of decades our research has taken its source of inspiration from *human institutions* [10,18,19] where the behaviour of interacting (human) agents is implicitly or explicitly governed. And if we take the time to reflect on human institutions, it seems clear that part of their success is due to the *values* of the participants (who return to the institution), the assumed values of other participants during interaction, and, critically, the values that are perceived by participants as being embodied by the institution itself. In other words, the governance system of human institutions embodies *values*.

Online Institutions (OIs) are our electronic equivalent of human institutions and they contain hybrid, multiple agents, whose interactions are governed, where agents can openly come and go, and which are situated in the real world. The problem of designing the governance for online institutions for hybrid (mixed) communities of computational and human agents so that they can be said to embody values is challenging for a number of reasons including how to know: (i) agents are who they say they are; (ii) the real motivation for agents joining the institution; (iii) that the effects of interactions don't lead to unexpected, unplanned and unwanted effects, (iv) that participants who behave badly will be appropriately punished; and (v) the explicit values the system is trying to uphold, and how what it does to ensure it is doing so.

The word "ethical" in the title is chosen to specifically refer to ways of measuring the extent to which a designed+deployed (d+d) system, and the behaviours it supports, can be said to implement a set of "values" that are explicitly articulated by the various stakeholders involved in the d+d system. Restricting the scope to OIs ensures that we can have clarity and focus on the issues involved in building ethical online systems of this kind, rather than tackling and claiming a solution for general AI systems. We build on twenty years of research in this area, ranging from formal definition and specification through to deployed and tested systems.

Even though our focus is on OIs, we believe that many of the challenges engineers face, and our proposals for addressing them, are relevant and applicable to a wider class of systems than just OIs. This focus enables us to build on research over the last decade on effective, ethical and conscientious design of OIs. We have set out to build a cradle to grave solution that provides accessible terminology to express the theoretical and practical issues involved and our responses to them.

Our goal is an intuitive, grounded, principled and practical approach, that builds on the WIT model [15], and a set of previously described over-arching Conscientious Design values [16], to show how to put both into applications, using methodological guidelines in the form of heuristics, for the actual embedding of values. The three research questions which outline the scope and ambition of our paper are: (i) how to take

human values and operationalise them into the workings of an OI? (ii) what are the relevant human values to embed in the operation of a given OI? and (iii) how to tell to what extent a designed system is consistent with those values?

2 The Conscientious Design story so far

2.1 The WIT Design Pattern

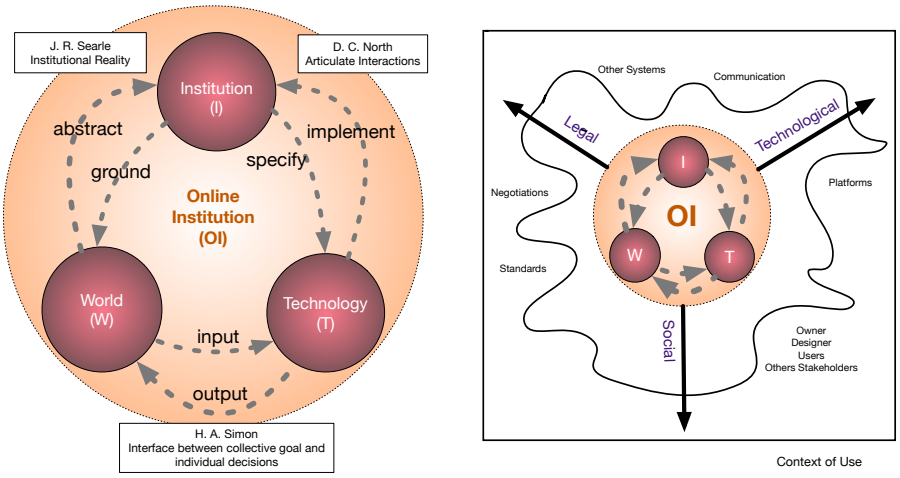
The $\mathcal{WIT}$ design pattern for online institutions has been developed over the last decade or so. The most recent description [17] is relatively high-level for a non-specialist audience, but at the same time the most developed contextualization to date, while earlier iterations at previous editions of COIN(E) [16,15] are more technical, and chronicle the evolution of the idea. We briefly review our current position based on [17], informed by hindsight of [15,14,16], for a technical audience and for the sake of grounding the terms used in the rest of the paper.

The purpose of the $\mathcal{WIT}$ design pattern is to support the process of building online institutions (OIs) (see Fig. 1a), which are open, regulated, online multiagent systems, characterised by an interaction space (referred to as the World) for human and software agents, Institutions that observe and regulate agent actions, and Technology that mediates the interaction. The one additional, critical characteristic is situatedness, which establishes that an online institution is anchored in and to the real world by legal, technological, and social constraints (Fig. 1b).

The first significant difference between earlier work, before [17], 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 [1,2] 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) [7,8,9] to provide the basis for the role of values in the design process, 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" or OI, which facilitates two abstractions: (i) the isolated OI in Fig. 1a, which enables the design of an OI to be considered from three different but related perspectives: W, the OI as seen from the world perspective; I, the institutional or governance perspective of the OI and 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" [12] in both directions and to anchor the online institutions with its physical world counterparts.

The isolated OI is the repository of the state of the world, as observed and interpreted by its constituent institutions. These components of the OI provide the affordances – what an actor perceives it can do with an institutional action – and determines which actions have institutional effect in a particular state of the world. The isolated OI should demonstrate *cohesiveness*, which is to say the three views work as intended, and *integrity*, which means it is a persistent, well-behaved online system. The situated OI, to be fit for its purpose, needs to be effective in the context of its use. Context will typically include its technological, legal, social and economic context, but will in general



(a) The Isolated OI, drawing on Searle [26], North [18] and Simon [27]

(b) The Situated OI

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

be determined by the intended usage. Clearly, for the situating to work, the situated OI should demonstrate *compatibility* with its context.We appreciate this reprise of WIT is short on detail, but hope it provides sufficient intuition to support the understanding of the example (Section 3), and the main body (Section 4) of the paper, where we look at the task of making values operational.

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* (CD) in recent work [17,16]. The CD framework rests upon three *value categories* which are titled: 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 [17]):

- **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. Robustness, resilience, accessibility, and security are all components.
- **Mindfulness**: refers to working through the range of impacts on human users so often over-looked. It is about engendering a wider awareness as we develop understandings of the way in which online systems can cause harm or improve well-being. Examples include for instance, data ownership and ease of access, and have 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 [25].

In our work 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 [29] on Ethically Aligned Design [17]. As meta-analyses of the multitude of frameworks show [5,13], many have overlapping definitions and principles. However, the CD value categories support more than one way of looking at each particular principle. This is a notable benefit of CD's principled approach.

One final remark here concerns the stakeholders. Stakeholders are all those affected by the development or the use of the system, however, usually not all stakeholders can affect the development. Those who are affected but are not part of the decision-making in the design of the system we will call *indirect* stakeholders —as is the usual term in value sensitive design. *Direct* stakeholders, therefore, are those whose values need to be primarily 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

Here we introduce a worked, fictitious example of an OI to support the understanding of the theoretical and practical concepts involved in the WIT-DP for ethical OIs. We call this OI Easyrider, and it is a system for buying and selling train tickets online. The four stakeholder groups (Owner, Engineer, Users and Indirect stakeholders) are as follows:

1. Owner: refers to the individual or organisation that commissions and run the OI. In this case the railway company that sets up Easyrider to sell tickets on line through travel agencies.
2. Engineer: refers to the entity responsible for ensuring the requirements of the owner are satisfied in the newly designed and deployed OI in a way which attracts users.
3. User(s): refers to *passengers* (who are human agents) that use Easyrider to buy and sell train tickets, and *travel agencies* (who are software agents) that buy tickets from the railway company to re-sell them to passengers.
4. Indirect stakeholders: in this case include, for example, 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 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 [23].

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 operation; (ii) the provision of a service through travel agencies that is profitable for these and attracts passengers; and (iii) an acknowledged positive impact because more persons travel in train instead of using less ecological means of transportation and the 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 and responsibly achieving a healthy cash-flow. Alongside, the OI promote 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 [22], 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.

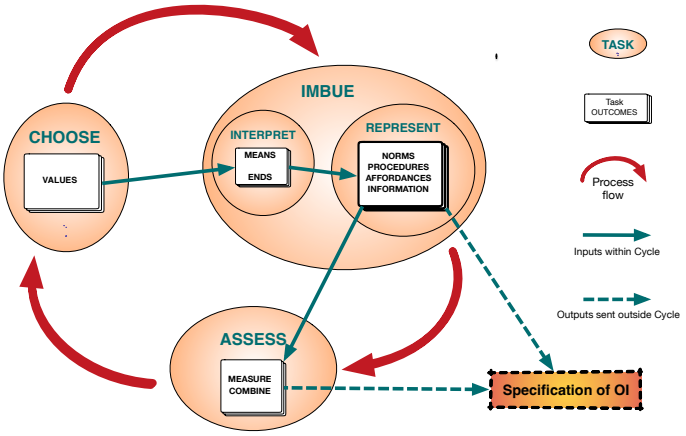


Fig. 2: The process for bringing values into the design of an OI

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 that (ii) facilitate the separation of design concerns and (iii) 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 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 co-design 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 then 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) [23]. 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 to be 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

In everyday language, we talk about values with terms that stand for, or "label", their true meaning [3]. These value labels are not arbitrary, they do reflect, however vaguely, some shared idea of what the value stands for and how one can tell if it is being upheld or not in given circumstances. The point of this task is to choose value labels that are not too vague, because their meaning needs to be interpreted and eventually represented in the OI (the next tasks in the operationalisation process). The trick is to decompose the problem (again) so that different concerns are identified and for this purpose we propose two heuristics.

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* makes some instrumental values relevant and others less so. For example, in Easyrider values related to e-commerce and transportation become relevant, while those associated with, say, health service do not. Second, the *role of stakeholders* influences the choice of values. However, regardless of the application domain, engineer values always reflect the goal of developing an OI that handles a particular collective activity (Table 1), 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.

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

A word on 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

Engineer
Ultimate goal: Design and build proficiently an OI
CD value categories:
Terminal and instrumental values:
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: guarantee cohesiveness and integrity and technological compatibility of the OI.
WIT contextualisation
Leading stakeholder for value operationalisation in (i) WIT contexts: $\mathcal{I}$ >$\mathcal{T}$; $\mathcal{T}$ > $\mathcal{I}$; $\mathcal{W}$-> $\mathcal{T}$ (ii) Integrity of isolated OI; (iii) Technological compatibility of situated OI

Table 1: Engineer's contextualisation of values for CD value categories and WIT pattern (regardless of OI's domain)

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; second, not all the nine contexts are equally important for all stakeholders, hence one can rank the degree of involvement 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 while table 2 (in the next section) illustrates the terminal and instrumental values of the other Easyrider stakeholders.

The second heuristic suggests how to proceed in order to identify relevant values.

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*.

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 (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 inhouse practices and systems (M), reliable support (M), ...*
Good customer relations *reliable and resilient 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: A schematic interpretation of some Easyrider values

The rationale is that each stakeholder has its own *ultimate goal* for their engagement in the OI and the definition and pursuit of these goals are determined by the stakeholder's *terminal values*. However, the way that such goals are in fact pursued within the OI are motivated by the instrumental values of the stakeholders. The CD value categories are needed to track that instrumental values do cover the three complementary concerns of every value-imbued OI.

We illustrate how this heuristic is applied in the Tables 1 and 3. In the first we show the engineer's choice of values for a generic application, while in the second we outline the specific choices for Easyrider of its owner and users.

Table 2 is a partial contextualisation of the terminal and instrumental values of the owner and the users of Easyrider (the engineer's values were discussed in Table 1). The point of the table is to use particular instrumental values (the ones that are underlined) to illustrate, in Sec. 4.3, the interpretation and representation of value labels.[5]

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

[5] The table shows instrumental values that are meant to be typical of one passenger and one travel agency; other individuals may interpret value labels differently. All values are labeled with the CD value categories they belong to.

the OI. This task of imbuing values in a system involves two efforts: *interpretation* and *representation* of values. These two processes are applied to each instrumental value label, all stakeholders are involved in this process but it is led by the stakeholder who holds that value.

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.

For the first purpose, 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 thresholds of satisfaction of the different stakeholders. 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)** provides the basis for measuring and combining values by identifying *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 autonomous 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, which is intended to influence their behaviour. Table 3 illustrates the interpretation and representation of some instrumental values (from 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 *transparent rules of operation* for stakeholders.

	Passenger	Users and owner	Owner	Owner
Values	Flexibility	(accountability, transparency)	support SDGs	adequate return on investment
Ends	Facilitate last-minute purchases	Proof of action	Promote the use of train to support SDG 7 , 9, 13	High occupancy of wagons
Means	Extend purchasing deadlines; install ticketing machines at station	Reports of relevant transactions	Marketing campaign	Attractive fares, ease of purchase, marketing
Observable	Number of tickets sold close to deadline; number of machine-issued tickets	List of reports of each type	Passenger and TA awareness of the good impact of trains	Occupancy rate
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 wagon
Representation	Norms and affordances	Procedures for issuing each report type.	Banners and messages, poll.	Procedure and physical action: add wagons when needed

Table 3: Examples of the imbuing of Easyrider direct stakeholders' instrumental values

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 outcomes that promote or protect a value. 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, and then delegate to the engineer the details of representation, specification and implementation that ensue.
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 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 through a customer satisfaction poll and satisfied as long as the aggregate opinion is positive or very positive.

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, denoted VAM_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 observing the outcomes that stand for the value and establishing a way to measure the satisfaction of the possible outcomes with respect to the preferences of the stakeholder.

For instance, in Easyrider, a travel agency recognises "accountability" as a responsibility value, which is being interpreted as "honouring deals". This instrumental value is interpreted, in particular, by guaranteeing that the travel agencies pay all their dues to the railway company and to other travel agencies. The means that the institution has implemented to maintain that value, is 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 measure accountability by the sum of fines it pays over the year and prefer to pay as little as possible.

While the example of accountability makes satisfaction look something like a utility function, in fact the only requirement is that satisfaction scores can be mapped to a preference relation. 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 (too many, 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 seated together, good). Satisfaction could be measured, for example with a pairwise preference combination of density vs seat configurations.

The last component of the VAM_s is to combine values using an aggregation function, in order to assess the extent to which the OI aligns with the stakeholder's values. The way the aggregation function is defined may take into account the priorities and trade-offs between values and other features like their urgency, associated costs or expected evolution.

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. It stands for the alignment of the OI to the values of that stakeholder.

Thus, the purpose of the aggregation function is two-fold: first to commit to a measure of satisfaction that reflects value priorities and trade-offs; and second use that measure of satisfaction to determine if the alignment is "good enough". Consequently, If the alignment is not good enough, the aggregation function, first, and the value assessment model in general can be used to pinpoint those values that are not properly embedded in the OI.

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.

That is, (i) start by revising and improving the aggregation functions, (ii) if that does not solve the problem, examine and improve the assessment functions for the three CD-value categories; (iii) if that is not enough, examine and revise the representation of a specific value. That is, start by revising its satisfaction thresholds, if this does not

improve the alignment, examine and revise the measurement, then the representation of the value, and finally means and ends. If none of the above offer adequate remediation, the conclusion is that the alignment is unsatisfactory for the given value and other values need to be examined and revised. At this point, (iv) the issue needs promoting to the next level of reflection [28] before revisiting the value operationalisation process from an earlier stage (structuring, choosing, imbuing, assessing).

A compromise can usually be reached by revising the aggregation function, simplifying value measurement and relaxing satisfaction thresholds.

The rationale is the following: the value alignment models of all the stakeholders are superficially identical, since all combine satisfaction with respect to the three CD value categories. However, each stakeholder has different terminal objectives and terminal values, so their aggregation functions are expected to differ not only in the way they combine satisfaction functions of the three CD value categories but also in the values that are taken into account for each category. Thus, improvement of value alignment tracks back through the operationalisation process to work out where to fix it.

5 Closing remarks

In this paper we propose heuristics to make the values of stakeholders operational in online institutions. These heuristics belong to a larger task of newly emerging methodological guidelines to support a principled approach to imbuing values in artificial autonomous intelligent systems. The fundamental assumption here is that the values to be embedded in any system are aligned with the ethical principles common to its stakeholders. We understand that this assumption requires that values are 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 unavoidable, we do not impose any further requirements to value theory beyond this objectivistic perspective.

Because of that neutrality with respect to value theory, the heuristics we propose remain neutral about the formalisms used for representation and for the assessment of values. However, we believe that for certain types of online institutions (and AIS in general) there are reasons to adopt specific interpretations of value means and ends that give grounds to more specific representation and assessment conventions, although we recognise they might not necessarily be unique. For example, we have argued in favour of some forms of consequentalism in the case of policy-making sandboxes, on the grounds of the methodological implications of using agent-based simulation for the assessment of policy outcomes in the policy-making cycle. However, even then, consequentalist views need not be made operational with conventional utility and welfare formalisms for value measurement and assessment [20,21].

Although individual artificial agents may be designed to assess values in the course of their ongoing autonomous behaviour, addressing the specifics of making values operational for the alignment of such behaviour is out of scope of this paper. The process of such operationalisation would be similar to that which we have outlined in our work on OIs, and it is likely most of our heuristics would still apply. However, there are specific aspects of the design process that would need to address the role of values in designing

autonomous architectures and behaviour. For instance, for an autonomous 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 [20].

We mention elsewhere [16] that one could apply the conscientious design approach to develop tools to prevent undesirable effects of existing third party software. The heuristics we propose in this paper may serve to diagnose alignment and, potentially, to identify some outcomes that need to be controlled. 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.

References

1. Alexander, C.: A pattern language: towns, buildings, construction. OUP (1977)
2. Alexander, C.: The timeless way of building, vol. 1. New York: OUP (1979)
3. Carroll, L.: Through the Looking-Glass, chap. VIII. MacMillan and Co. (1871), Haddocks' Eyes: https://en.wikipedia.org/wiki/Haddocks%27_Eyes
4. Edwards Deming, W.: Quality, productivity, and competitive position. MIT Press (1982), see 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. 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)
6. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al.: AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines **28**(4), 689–707 (2018)
7. Friedman, B.: Value-sensitive design. Interactions **3**(6), 16–23 (1996)
8. Friedman, B.: The ethics of system design. Computers, Ethics and Society pp. 55–63 (2003)
9. 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)
10. Harré, R., Secord, P.: The Explanation of Social Behaviour. Blackwells (1972)
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. Jones, A.J.I., Sergot, M.: A Formal Characterisation of Institutionalised Power. Logic Journal of the IGPL **4**(3), 427–443 (06 1996)
13. 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)
14. 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.
15. 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)

16. 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)
17. Noriega, P., Verhagen, H., Padget, J., d'Inverno, M.: Ethical online AI systems through conscientious design. IEEE Internet Computing **25**(6), 58–64 (2021)
18. North, D.: Institutions, Institutional Change and Economic Performance. CUP (1991)
19. Ostrom, E.: Governing the Commons. The Evolutions of Institutions for Collective Action. Cambridge University Press, Cambridge (1990)
20. 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)
21. 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)
22. van de Poel, I.: Embedding values in artificial intelligence (AI) systems. Minds and Machines **30**(3), 385–409 (2020)
23. Rokeach, M.: The nature of human values. Free press (1973)
24. Russell, S.: Living with artificial intelligence (Dec 2021), https://www.bbc.co.uk/programmes/b00729d9/episodes/downloads
25. Schwartz, S.H.: An overview of the Schwartz theory of basic values. Online readings in Psychology and Culture **2**(1), 11 (2012)
26. Searle, J.R.: The Construction of Social Reality. Allen Lane, The Penguin Press (1995)
27. Simon, H.A.: Models of man; social and rational. Wiley (1957)
28. Smith, B.C.: Procedural reflection in programming languages. Ph.D. thesis, Massachusetts Institute of Technology (1982), http://hdl.handle.net/1721.1/15961
29. 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