Ethical Online AI Systems through Conscientious Design

Pablo Noriega
IIIA-CSIC

Harko Verhagen
Stockholm University

Julian Padget
University of Bath

Mark d'Inverno
Goldsmiths, University of London

***Abstract*—The increased interplay between humans and artificial agents especially in online environments asks for reliable ways to impose reliable governing principles for these interactions. Grounding these principles on a set of interrelated values to ensure alignment of the stakeholder values and the behaviour in the hybrid online interaction space is the proposal made in this paper. A tripartite model of the interaction space helps developing the mechanisms needed, resulting in online institutions. These parts together form conscientious design.**

Introduction

We are entering a time where humans will increasingly interact with artificial intelligent systems (AIS) in online environments, and this interaction will take place in ever more sophisticated ways. So not only will there be more online communities of human and computational agents, but the artificial systems will have increasing levels of sophistication in the way they interact and greater autonomy over their decision making. This suggests a pressing need to look at approaches to the design of such systems so that there is confidence they are places we would wish to inhabit.

Greater autonomy means greater potential to impact on the social and psychological states of human participants. This potential raises new concerns about how we can protect their well-being when these increasingly sophisticated computational agents might now be as untrustworthy and malevolent as they are incompetent.

Understanding the ways in which we can protect the well-being of humans should provoke us to consider the ethical responsibilities when designing AIS. We want to be able to harness the autonomy of agents in hybrid communities but do so in ways which are safe for human users. Yet whilst engineering ethical considerations into AIS has often been spoken about, current practice is patchy at best. And even if ethics are considered, ensuring that the design and implementation of a system which supports human and artificial agents does not harm humans is rarely undertaken in any systematic or principled way.

Part of the reason for this is that setting out to respond to questions such as: what does it mean to do the "right" thing?, how can it be known with any degree of certainty that a new AIS system will support the "right" thing?, and when is enough "enough" in terms of what needs to be thought about?, are not clear.

Furthermore, the risks of getting it "wrong",

and a new system causing harm, are hard to assess too. Not only because all kinds of unplanned behaviours and impacts could emerge, but also because of a lack of documented experience in addressing ethical issues in AIS design. Because considering these factors together is so hard, it might lead to ignoring the issue altogether or hoping that basic common sense will be enough and that things can be worked out on the fly.

In response to these concerns, we have developed the notion of Conscientious Design (CD) which aims to be a systematic and practical approach to support practitioners in the ethical design of AISs. Right at the offset, we want to be clear that this is not just "yet another methodology" for designing ethical systems, and that this is true for two key reasons. First, it is an approach that builds on the principles and practices of well-established practices in value-sensitive design (VSD) [1], Alexander's "habitable spaces" [2], and Deming's total quality management (TQM) [3]. Second, it provides a way of using familiar agile concepts to imbue an AIS from an ethical standpoint. Additionally, it puts human and software (AI) participants in control of co-evolution of the online spaces they jointly inhabit.

The participants in Alexander's habitable spaces referenced above are physically constrained; those in online spaces are constrained too, but in different ways. First, online spaces are constrained by the platform itself and what it allows. These are known as platform-provided affordances [4] (e.g., "buy", "like", "ban"). Actions not allowed by the platform simply cannot take place. Second, actions of one participant are constrained by the normative expectations that the other participants have of what is considered acceptable and unacceptable behaviour [5], [6] (e.g., spamming, helping, ignoring), where non-compliance may lead to sanctions against the acting agent. Note that whilst there might be a degree of homogeneity in the normative expectations of others, they do not need to be identical. These two categories of constraint are perhaps most easily understood through our individual experiences of using on-line platforms (e.g., shopping, social networks).

For some years now we have been researching a particular subclass of AIS called online institutions (OIs) [7], [8], [9], [10], [11]. OIs contain policies that facilitate the governance of participant activity, either through what a participant is allowed to do in certain circumstances or what a participant may choose (not) to do for the sake of any social consequences. Online institutions embody both affordances and norms, thus interpreting Alexander's "Timeless way of building" for the social – often commercial – spaces in which we participate on the Internet. Furthermore, OIs (as with all AISs) being a software construct, have an intrinsic adaptability and resilience, which means that they can in theory support Deming's evolutionary approach to the achievement of quality over time, founded on VSD's value principles. Furthermore, we believe that by considering online institutions we can most effectively map out the principles and building blocks of conscientious design. From this basis we can then establish what is generalisable to other categories of AISs.

One of our guiding principles is to set out the ideas underlying conscientious design in a way that can bring researchers and practitioners together from different disciplines. This means marrying the inherent complexity that comes with considering issues of affordances, governance and autonomy with a clarity of setting out the issues, and the way in which any design process needs to address them. We believe that the time to do so is now, because the increasing prevalence of AISs with greater artificial autonomy we mentioned above, brings with it a very significant risk to societal well-being if we do not.

When introducing a new approach to design it is necessary to situate and contextualise the ideas it contains through current examples of systems that have not used the approach. It is in this way that we aim to show how CD offers developers a lens through which to review existing systems for adherence to CD principles, and thereby to values and ethics. Moreover, it supports a practical way to learn what we need by looking at what is missing from those designs, which in due course can lead to more principled approaches to the design and implementation of online institutions.

Conscientious Design

Stakeholders in VSD identify the values that characterise the most important properties of the

system that they wish to build. For example, the Estonian e-justice system coalesced around transparency, integrity and security through a process involving relevant ministries, actors in the justice sector, and citizens. Other systems coalesce around other values which is what VSD explicitly sets out to support. This process creates two challenges: how to identify the (small) set of core values and to which value or values to associate different aspects of the design, without connecting everything to everything. CD builds on VSD by providing a frame of reference for the stakeholders' values by proposing three value sorts: thoroughness, mindfulness and responsibility. We will provide general characterisations of these headings and claim that when grounded by the chosen values and a system to build, can help co-designers debate how those values contribute to the overall design. It will help determine where there is mutual reinforcement, where there are conflicts, and where elements may be missing.

Stakeholders in VSD are presented with a simple ethical framework: first consider what is right, and secondly what is good [12], which hints at an hierarchy of values and debates over which values are right and which are good. CD nuances this debate by offering sorts of values that provide a frame in which to argue about the "how and why" of the network of stakeholder values and the relationships with the sorts, rather than whether one value is more important than another. The three value sorts are not arbitrary, but derive from global studies of values across cultures [13], [14], [15], aiming to capture the centrality of three kinds of interrelated values:

- **Thoroughness**: this refers to conventional technological values that promote the technical quality of the system. In any *(stand-alone) system* values include completeness and correctness of the specification and implementation, reliability and efficiency of the run-time version of the system, robustness, resilience, accessibility and security. For any *situated system*, these include technological compatibility, security against intrusion, and integrity of data and communication.
- **Mindfulness**:
  We have chosen this word carefully to respond to the considerations about impact on human users that are so often over-looked. In its characterisation mindfulness includes building a wider awareness of what is happening around us in order to make the right choices, in line with Deming's principles. Examples of values in this category concern *data ownership* (privacy, data agency, usage traces), and *well-being* (accessibility, respect of user's attention).
- **Responsibility**: these are values that address the anchoring of the system (towards the owner, the users, and any external stakeholders). Here, we can think of the effects of the system on the context in which it is situated (liability, accountability), and how that context may affect intended users and external stakeholders (legitimacy, user protection, no hidden agency).

Like any research, the idea of conscientious design builds on a range of existing work but also makes significant contributions. One contribution is how it supports current initiatives from the EU and IEEE on building AIS. The first are the guidelines which come from the EU's High Level Expert Group on AI for the development of trustworthy AI [16], the other is the IEEE's vision for Ethically Aligned Design [17]. CD is not an alternative to these, but rather a way to make each of these work, indeed its development goes back to well before these two projects. Indeed, these projects underline the timeliness of CD. In table 1 we illustrate how CD values relate to the EU and IEEE principles respectively, based on the keywords used in the documents in which they are described. For instance, the EU Guidelines have under the ethical principle of explicability the following example measures: "traceability, auditability and transparent communication on system capabilities" [16]. These belong to the CD value of responsibility, in that they describe the anchoring of the system. As an example of the mapping of the IEEE ethical design principles, consider competence. This addresses safe and effective operation [17], i.e. it belongs to the CD value of thoroughness, with its focus on the technical quality of the system.

Online institutions and the WIT pattern

The next step in the presentation of CD is to provide a framework for the operationalization of

		Thoroughness	Mindfulness	Responsibility
EU HLEG Guidelines for Trustworthy AI Ethical Principles	**Human autonomy**			✓
	Prevention of harm	✓	✓	✓
	Fairness		✓	✓
	Explicability			✓
General Principles of Ethically Aligned Design for Autonomous/Intelligent Systems	**Human rights**			✓
	Well-being		✓	
	Data agency		✓	✓
	Effectiveness	✓		
	Transparency		✓	✓
	Accountability			✓
	Awareness of Misuse		✓	✓
	Competence	✓		

Table 1: Mapping EU and IEEE principles onto CD values

the CD values in the construction of sociotechnical systems containing AIS. The sociotechnical systems approach is a well-established methodology to analyse (in the design phase or in the deployment phase) intertwined systems consisting of human agents, technological artefacts, and institutional rules. However, in the AIS case, the autonomy and adaptability of the artificial agents surpass traditional technological systems. Instead, they can be seen as sociotechnical systems themselves, thus asking for special kinds of institutional policies covering the AIS [18]. This forms the core of the online institutions introduced in the previous section.

We made it clear in the previous section that online institutions play a critical role as repositories of policies for software to interpret. For software components with the requisite reasoning capabilities – which could range from "do what the OI tells you to do" to being able to assess the pros and cons of compliance with the OI guidance – this provides a form of late behavioural binding that can be viewed as the multiagent systems complement of object composition (aka black-box inheritance): change the OI to change the behaviour. The OI is in effect the conduit for the expression of human intentions and expectations in the sociotechnical system, to provide a form of requirements at runtime [19].

The CD values aim to help designers in debating the why and what of the system, but the translation from what to how needs equally careful handling in order to maintain the separations of concern established in the earlier stage. As with Alexander's blueprints, the objective is not to provide an answer, but a way to think about the answer and arrive at a (different) appropriate solution every time. For this purpose we propose the World-Institution-Technology pattern (see Fig. 1), as the sociotechnical systems complement of object-oriented programming's Model-View-Controller, where the world (W) is a collection of social spaces, that are sub-contexts of the real world, institutions (I) are the policy frameworks into which the values that characterise the system are imbued, and the technological space (T) where online interactions are processed according to software representations of the institutional conventions.

Online institutions in CD are the glue that binds such systems together, mirroring the functions of conventional social and economic institutions [20], [5], [21]. This class of sociotechnical systems is formally defined in [9], [22] and is a refinement of other abstractions of systems for social coordination and artificial or electronic institutions [11], [23], [8]. Informally, an OI provides technological support for human and software agents to interact online with each other, and establishes the policy – the "rules of the game" – that governs those interactions. The terms of the policy determine what fragment of the real world is relevant, what events and actions that take place in the world are recognised by the institution and what their effects in the institution are, and *vice versa*. For this purpose,

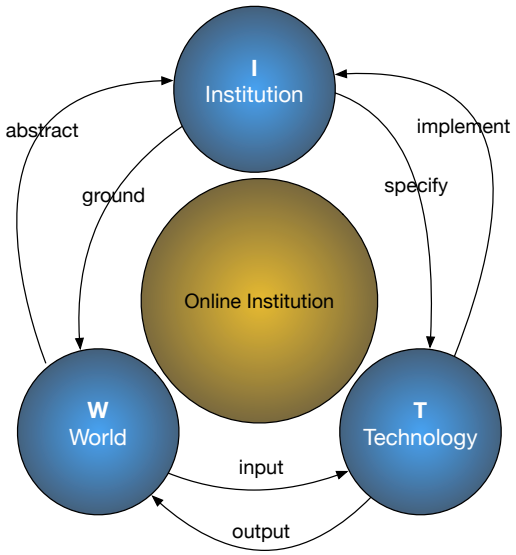

Figure 1: World + Institution + Technology perspectives for online institutions and their interrelationships

an OI (i) maintains an *institutional state* that is accessible to all the active participants and (ii) depending on the agent's role, the action, and the current institutional state, may recognise an action as correct and update the institutional state – signalling that a pick-up has occurred via the Uber app – recognise an action as incorrect and signal non-compliance – signalling a driver missed a pick-up – or ignore the action and leave the institutional state unchanged.

To expand on the WIT model in Figure 1, we now look in more detail at the relationships between the parts that make up the model:

1) W↔I: intuitively, I is an abstraction of the relevant sub-context in W, that captures "just enough" of the real world dynamics – the actions and events that can occur that matter for the sub-context, like movement, or picking up or dropping items in a game – and an institutional model that represents the policy that applies to those recognised actions and events. In the other direction, institutional changes need grounding to have consequences in the social (world) context, such as a passenger rating affecting driver selection in Uber;
2) I↔T: the abstraction in I provides the specification for what has to happen in T, telling the developer what function the technology space should deliver, while the relationship in the other direction documents how the technology space does what I specifies;
3) W↔T: lastly, it is the relationship between W and T that enables the participants of the social (world) context to interact, by whatever interfaces are appropriate (phones, game handsets, VR, sensors of various kinds) providing inputs to the OIs (actions and events) and receiving output (institutional interpretations and institutional consequences of those actions and events).

Moreover, the WIT pattern not only reveals different interrelated aspects of an OI as a *standalone* system, but it also helps to understand what to take into account when examining the OI as a system that is *situated* in its (evolving) working context; in particular to analyse its compatibility with the legal, social and technological environments in which it is deployed.

The key takeaways from this part of the CD story are that (i) for the kinds of sociotechnical system at which CD is aimed, online institutions are a means to provide a flexible, transparent and interpretable – to humans and software – representation of policies, and (ii) the WIT pattern offers a separation of concerns between the participants, and the policies that govern and the technology that mediates their interaction.

Using values in online institutions

Values are powerful and practical devices to imbue ethical behaviour in AIS. In general, values serve two main purposes: to assess the "worthiness" of a state of affairs and to decide what is the "right" action to take [13]. Thus, within OIs, values will be reflected in the actions that may be accomplished within the institution and their effects. So, institutional governance should promote or require actions and effects that align with stakeholders values and prevent or discourage those that do not.

The challenge is to make values operational. This involves three processes: (i) one needs to *interpret* the meaning of each value, (ii) choose means to *implement* values in the OI or the artificial agent and (iii) be able to *assess* whether they are being attained and to what degree. For sake of brevity we mention mostly the positive aspect

of "desirable" behaviour but what we propose below also applies to the handling of undesirable behaviour.

The *interpretation* of a given value consists of identifying behaviours and outcomes that are characteristic of that value so that these are encouraged or guaranteed to happen. Values, at face value, represent good intentions that need to be made operational. Thus the three CD values need to be instantiated with different concrete values that allow for a refinement of its interpretation, implementation, and assessment. One needs to take into account that interpretations of the same value may vary depending on the context in which the behaviour is to be observed, the perspective of the stakeholder who observes it, and the moment when the value is assessed. The WIT pattern facilitates this analysis, as we discuss at the end of this section. With these provisos in mind, there are two approaches for defining the meaning of a value. One, is to produce an explicit description of behaviours that uphold the value (or demerit). Another way of interpreting a value is to choose a set of indicators – observable parameters in the state of the system – that reflect support for the value (or its demotion).

The *Implementation* of a value can be achieved by focusing on the behaviours and outcomes that are aligned with the value. Note that values may either be expressed through policies to govern the collective behaviour inside an OI, or that of an autonomous agent. The first case implements values by way of what van de Poel [18] calls technical norms. The second case is similar but involves the additional components associated with value assessment and decision-making on which we comment below.

There are three basic tools for implementing values in an OI:

1) *Hard-wiring constraints and procedures* that implement specific behaviour and indicators associated with the interpretation of values. This presupposes the choice of the relevant entities in W that will provide the basis for the institutional model and its implementation. This hard-wiring needs to adapt to the evolution of an OI. For instance, in online multiplayer games such as League of Legends, the base capacities and skills of the characters the players can choose from are given, as are the ways in which these can be extended during game-play.
2) *As explicit policies that are part of the institutional model*. These may comprise (i) *functional* norms that specify the preconditions and the effects of admissible actions; and can thus be easily linked with indicators. These norms may include incentives and disincentives as well as assignment and removal of entitlements, obligations and permissions to individual agents; or (ii) *procedural* norms that define how to perform and implement a specific behaviour that interprets a value. Note that these norms are enforced by the OI and this enforcement may be strict (regimented) or not (enforced according to some institutional conventions, such as actions taken by other participants). In *Uber*, a "fairness" norm assigns a rider the closest available car but gives preference to cars with higher client satisfaction ratings.
3) *Influencing decision-models of participants* by providing additional information or arguments that may promote a change of decisions. In the case of of online games, such as League of Legends, the problem of toxic gaming and inappropriate language between temporary teammates is detrimental for the enjoyment gaming is supposed to give. In League of Legends, at first a sanctioning strategy was chosen – initially using selected human players as a jury to judge complaints [24], later on replaced by an automated sanctioning system which was criticised, amongst other reasons, for not being transparent [25]. In its latest incarnation, a positive reward system has been put in place as an honour system in which team mates can give each other positive feedback. How this feedback is represented in the game (a badge with a numerical value) and what it may result in (extra in-game rewards) has changed over time but an overall critique remains to this system as well: it is the game company who decides what is and what is not transgressing the "honour rules of the game" [26], i.e. not all stakeholders are part of the discussion on how to assess the fulfilment of the value of "fun".

Value implementation for an individual agent is quite similar: *hard-wiring* is achieved with a repository of standard behaviours and means to choose the appropriate one in a given circumstance and *norms* may embedded in the decision-model of the autonomous agent. The main difference is that for modelling an individual agent one needs to implement the choice function (what action to take) as part of its decision model. Depending on the agent architecture this may be more or less complex.Note that ofttimes the designer of an OI (and the individual agent) needs to assess several values simultaneously. For that aggregate assessment one needs to make their individual assessments somehow commensurate.

Assessment of value attainment can be achieved by measuring indicators or by determining whether the intended behaviour is actually performed. According to [18], this is most easily done by checking the explicit implemented AIS norms and normative reasoning. An important benefit of WIT is that it allows separation of concerns at design and validation stages of an online institution in two ways:

1) *For the standalone system*, WIT helps differentiate and tailor the requirements associated with the interpretation of values in Figure 1. For example, *thoroughness* requires among other things, in W↔I, good alignment between the value indicators and feasible actions; in I↔T, a sound treatment of the evolution of permissions; and in W↔T a solid alignment between norm enforcement and interfaces.
2) *For the situated system* WIT supports discernment of the requirements for legal (W↔I), technological (I↔T), and social compatibility (W↔T), needed for the effective use of the system. For example, *mindfulness* entails, respectively, contracts that establish the limitation of liability of agents' actions; proper allocation of individual commitments and the transactions supported with the system; and culturally adequate interface conventions. Similarly *responsibility* requires, for example, that the system protects, advises and compensates agents regarding any breach of privacy or unwarranted financial costs.

Figure 2 shows how the WIT pattern facilitates separation of concerns at design time. In particular, it shows how value definition and assessment depends on the context where they are to be embedded and how different stakeholders are more or less influential in each of type of context at design time. Thus, builders play prevalent roles mostly in regards to values of thoroughness, while users in mindfulness and owners in responsibility but in looking at the WIT contexts, owners lead in the I node, since they define how the OI should function; while users and builders are, respectively, prevalent in W (what purpose the OI serves) and T (how it is implemented).

Furthermore, the meaning and assessment of values singly and severally depends not only on the perspective of a stakeholder but also on the time (e.g., design or run-time) of the assessment. Once the system is to be released and working, ex-post assessment is also stakeholder dependent. In rough terms, as Figure 2 illustrates, builders assess the system, in terms of their individual values: with respect to the fitness of the system to the needs and preferences of users, and to its fitness with the owner's business model. Users with respect to the effectiveness and convenience towards the satisfaction of their own goals. Owners assess the success of the system for, both, the collective goal and the business objectives.

We put all these considerations together into the notion of *Value Assessment Framework* which is nothing more than the explicit enumeration for each stakeholder of the following items: (i) the values that are relevant for the stakeholder in the specific assessment context; (ii) for each value its corresponding interpretation and assessment mechanism; (iii) the aggregation function for the set of values (and other conflict-resolution devices if needed).

This approach to assessment contrasts with the approach to value assessment chosen in [16], which is in the form of a checklist of questions, many of which can be answered with yes or no. In our approach, such a potentially superficial set of answers is replaced by requiring the designer and stakeholders to analyse in-depth how and why the overarching conscientious design values are supported via the way they are imbued in the system.

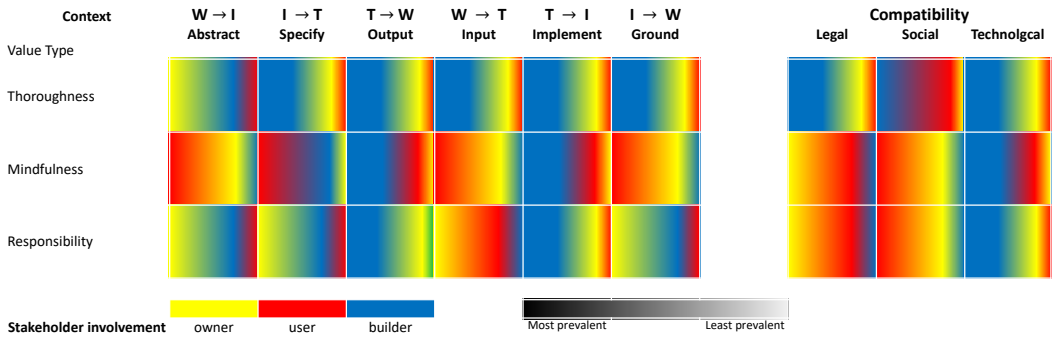

Figure 2: Value interpretation and assessment varies with stakeholders and context of application. The WIT pattern facilitates the separation of concerns so that each stakeholder may be involved with a different degree of intensity in different contexts. The proportion of colour for each stakeholder indicates their likely degree of engagement with each value and each WIT relationship (left), and similarly for value and situational factors (right).

In our description of the value assessment framework we do not commit to any particular value interpretation schema, nor to any specific aggregation and conflict resolution device. This is not laxness, but rather to allow conscientious designers to be conscientious in making appropriate, defensible choices.

Concluding Remarks

As we look at the world around us, we see a diverse range of popular sociotechnical systems with mixed communities of interacting human and software participants in action. Most of these have been built, and are being used, without recognising that a new kind of approach to design is needed to protect the well-being of the human users. We cited several examples of such systems in different domains with high use earlier in this paper. In each of these domains we can find examples of platforms where (i) human interaction with artificial agents is mediated, and (ii) artificial agents interact with humans in the same online space and so directly impact human users.

In this article we have made reference to the WIT model for looking at *online institutions* which are a subset of AISs, where governance is explicitly represented. We have made the case of how we need to separate the concerns of online institutions into the view of the world, the institution, and the technology which supports the interface between the real world and the institution.

As we have demonstrated, existing systems normally only address the W→T relationship, and offer little in the way of either transparency or governance, which are essentially intertwined. Typically these are not even considered by the designers, and even if they are, they are hidden because they have used conventional design and implementation techniques that miss the new complexity that arises through issues of governance and artificial autonomy in hybrid communities.

The purpose of the CD approach is to support developers of ethical hybrid online systems in two ways. First, to provide a blueprint for the construction of online systems that we would be happy to inhabit achieved through the separation of world, institution, and technological concerns so as to facilitate the design of online institutions. Second, to enable the design of explicit, transparent governance mechanisms that contain mutually comprehensible representations of human-authored policies to describe what all participants may do under what circumstances.

The purpose of these two considerations is to bring ethical considerations systematically into the design process. They enable designers and other stakeholders to explicitly introduce their own values into the design of ethical AISs together with a balanced focus on affordances and norms. The CD approach also enables the system to adapt transparently as the changing needs and value priorities of users and other stakeholders change over time. One benefit of this, is that for

long-lived online systems, participants may self-organise with more explicit certainty over any potential harmful impact on users.

In this way, the entire stakeholder population acquires control of the continuous development and social utility of the system. This is different from the mainstream model where it is those external to the system who have the power to make and to impose change in support of their own (possibly hidden) goals, which may not be aligned with those who make the system live and create value.

In closing we would like to enumerate why the CD approach is significant. In doing so we hope to start to build a community of researchers and practitioners interested in the conscientious design approach.

1) CD is **relevant** because it provides an intuitive way to operationalize the principles set out in the trustworthy AI [16] and the ethically aligned design [17] guidelines.
2) CD is **methodology re-use**: it extracts elements from value-sensitive design, design patterns, and process quality to apply known thinking from agile development to target a class of internet-based systems; the "methods" already exist, CD reorients them for AIS.
3) CD is **timely** because we are still in the relatively early stages of the construction of sociotechnical systems, where although we run the risk of being impaled on the horns of Collingridge's dilemma [27][1], it is not yet too late to do something about it.
4) CD is **practical**: value imbuing is not a trivial process but our experience shows that it can be tackled with a principled strategy that interprets conscientious values in the various relevant contexts (stakeholders, stand-alone, situated) and uses adequate devices for making them operational (value interpretation, instrumentation, measurement, aggregation).
5) CD is **malleable**. Conscientious design assumes an ongoing implementation process involving the stakeholders from the start. In the "agile" ethos of conscientious design, values are not set in stone; with CD, they are identified and fit (ex-ante) to the specific context and are then assessed and progressively adapted ex-post.
6) CD **promotes and protects**. Conscientious design has a dual use. It is conceived as a systematic strategy for the design and evolution of new online institutions but it also can be applied to analyse and identify improvements – as modifications or add-ons – for refactoring conscientious values into existing systems.

We hope this work can be the start of an interdisciplinary community of researchers and practitioners who can together develop rigorous descriptions of CD components, document use cases where ethical considerations have been built into the design process that can be used for good practice, apply CD in AIS development from day one, and build a framework where values can be represented explicitly. We welcome anyone keen to join us in taking on these challenges.

[1] https://en.wikipedia.org/wiki/Collingridge_dilemma [Accessed 2021-06-28]

REFERENCES

1. B. Friedman, D. G. Hendry, and A. Borning, "A survey of value sensitive design methods," *Foundations and Trends in Human-Computer Interaction*, vol. 11, no. 2, pp. 63–125, 2017.
2. C. Alexander, *The timeless way of building*, vol. 1. New York: Oxford University Press, 1979.
3. W. Edwards Deming, *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.
4. J. J. Gibson, "The theory of affordances. the ecological approach to visual perception," 1979.
5. D. North, *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, 1991.
6. E. Ostrom, *Governing the Commons. The Evolutions of Institutions for Collective Action*. Cambridge: Cambridge University Press, 1990.
7. J.-A. Rodríguez, P. Noriega, C. Sierra, and J. Padget, "FM96.5 A Java-based Electronic Auction House," in *Proceedings of 2nd Conference on Practical Applications of Intelligent Agents and MultiAgent Technology (PAAM'97)*, (London, UK), pp. 207–224, 4 1997. ISBN 0-9525554-6-8.
8. M. d'Inverno, M. Luck, P. Noriega, J. A. Rodriguez-Aguilar, and C. Sierra, "Communicating open systems," *Artificial Intelligence*, vol. 186, no. 0, pp. 38 – 94, 2012.

9. P. Noriega, J. Padget, H. Verhagen, and M. d'Inverno, "Towards a framework for socio-cognitive technical systems," in *Coordination, Organizations, Institutions, and Norms in Agent Systems X*, vol. 9372 of *Lecture Notes in Computer Science*, pp. 164–181, Berlin / Heidelberg: Springer International Publishing, 2015.
10. P. Noriega, H. Verhagen, M. d'Inverno, and J. Padget, "A manifesto for conscientious design of hybrid online social systems," in *Coordination, Organizations, Institutions, and Norms in Agent Systems XII - COIN 2016 International Workshops, COIN@AAMAS, Singapore, Singapore, May 9, 2016, COIN@ECAI, The Hague, The Netherlands, August 30, 2016, Revised Selected Papers*, pp. 60–78, 2016.
11. H. Aldewereld, O. Boissier, V. Dignum, P. Noriega, and J. Padget, eds., *Social Coordination Frameworks for Social Technical Systems*, vol. 30 of *Law, Governance and Technology Series*. Springer International Publishing, July 2016. DOI: 10.1007/978-3-319-33570-4, ISBN: 978-3-319-33568-1 (hardcover), ISBN: 978-3-319-33570-4 (ebook).
12. B. Friedman, "The ethics of system design," *Computers, Ethics and Society*, pp. 55–63, 2003.
13. S. H. Schwartz, "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries," in *Advances in experimental social psychology*, vol. 25, pp. 1–65, Elsevier, 1992.
14. G. Hofstede, *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications, 2nd ed., 2003.
15. R. Inglehart, *Human beliefs and values: A cross-cultural sourcebook based on the 1999-2002 values surveys*. Siglo XXI, 2004.
16. High-Level Expert Group on AI (AI HLEG), "Ethics guidelines for trustworthy AI," Apr. 2019.
17. 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.
18. I. van de Poel, "Embedding values in artificial intelligence (AI) systems," *Minds and Machines*, vol. 30, no. 3, pp. 385–409, 2020.
19. J. Padget, E. E. Elakehal, K. Satoh, and F. Ishikawa, "On requirements representation and reasoning using answer set programming," in *IEEE 1st International Workshop on Artificial Intelligence for Requirements Engineering, AIRE 2014, 26 August, 2014, Karlskrona, Sweden* (N. Bencomo, J. Cleland-Huang, J. Guo, and R. Harrison, eds.), pp. 35–42, IEEE, 2014.
20. J. R. Searle, "What is an institution?," *Journal of Institutional Economics*, vol. 1, no. 01, pp. 1–22, 2005.
21. H. A. Simon, *The Sciences of the Artificial*. MIT Press, third ed., 1996.
22. P. Noriega, J. Padget, and H. Verhagen, "Anchoring online institutions," in *Anchoring Institutions. Democracy and Regulations in a Global and Semi-automated World* (P. Casanovas and J. J. Moreso, eds.), Law, Governance and Technology Series, Springer-Verlag GmbH, 2021. In press.
23. N. Fornara, H. L. Cardoso, P. Noriega, E. Oliveira, C. Tampitsikas, and M. I. Schumacher, "Modelling agent institutions," in *Agreement Technologies* (S. Ossowski, ed.), no. 8 in Law, Governance and Technology Series, ch. 18, pp. 277–307, Springer-Verlag GmdH, 2013.
24. M. Johansson, H. Verhagen, and Y. Kou, "I am being watched by the tribunal: Trust and control in multiplayer online battle arena games," in *Proceedings of of the 10th International Conference on the Foundations of Digital Games (FDG 2015)*, 2015.
25. Y. Kou and X. Gui, "When code governs community," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.
26. S. Tomkinson and B. van den Ende, "'thank you for your compliance': Overwatch as a disciplinary system," *Games and Culture*, p. 15554120211026257, 2021.
27. D. Collingridge, *Social Control of Technology*. Open University Press, 1981. ISBN-13: 9780335100316.