The Role of Engineers in Harmonising Human Values for AI Systems Design

Steven Umbrello (steven.umbrello@unito.it)
Institute for Ethics and Emerging Technologies; University of Turin https://orcid.org/0000-0003-2594-6313

Research Article

Keywords: value sensitive design, VSD, systems thinking, systems engineering, human values

Posted Date: September 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-709596/v1

Version of Record: A version of this preprint was published at Journal of Responsible Technology on April 1st, 2022. See the published version at https://doi.org/10.1016/j.jrt.2022.100031.

Abstract

Most engineers work within social structures governing and governed by a set of values that primarily emphasise economic concerns. The majority of innovations derive from these loci. Given the effects of these innovations on various communities, it is imperative that the values they embody are aligned with those societies. Like other transformative technologies, artificial intelligence systems can be designed by a single organisation but be diffused globally, demonstrating impacts over time. This paper argues that in order to design for this broad stakeholder group, engineers must adopt a systems thinking approach that allows them to understand the sociotechnicity of artificial intelligence systems across sociocultural domains. It claims that value sensitive design, and *envisioning cards* in particular, provides a solid first step towards helping designers harmonise human values, understood across spatiotemporal boundaries, with economic values, rather than the former coming at the opportunity cost of the latter.

1. Introduction

Computing technologies are becoming ever more pervasive in contemporary societies, to the point that discrete technologies are now inseparable from the understanding of social structures and institutions. In many ways, modern computing technologies manifest the long-held claim that technologies are sociotechnical (Verbeek, 2008). They cannot be understood as separate instruments, but rather as co-constituting and co-constructed by social forces. The sociotechnical worlds in which humans are continually immersed, now made even more manifest due to the SARS-CoV-2 restrictions, encourage a more granular evaluation of how they are designed and their corresponding impact on human values.

Artificial intelligence (AI) systems do not emerge *ex nihilo*, but are constructed and designed entities. In his famous article *Do artifacts have politics?* Langdon Winner (2003) demonstrated how the architect Robert Moses designed the overpasses across Long Island, NY to be intentionally low hanging to prevent already poor and segregated minority groups from accessing his prized beaches. Winner showed how even simple technologies, like the Long Island bridges, could support or constrain specific human values, whether designed intentionally or not. Given their current impact on quotidian human life, AI systems already, and will continue to, implicate a wide array of values (or disvalues) (van de Poel, 2020). Because they are designed artefacts, closer analysis is warranted to examine the nexus from which these innovations emerge, i.e., the design domains in which designers find themselves constructing these systems.

Engineers and designers work within social structures governing and governed by a set of values that primarily emphasise economic concerns. The majority of innovations derive from these loci. Given the effects of these innovations on society, the values they embody must be aligned with the stakeholders of those societies. Like other transformative technologies, AI systems can be designed by a single organisation or a consortium, but they are nonetheless distributed globally. Whereas previous research has focused on the socioethical impacts of AI (Bostrom, 2012; Floridi et al., 2018; Stahl, 2004), as well as how best to govern AI (Armstrong et al., 2012; Umbrello et al., 2021), this paper is the first to argue that (1) in order to design for this broad stakeholder group (i.e., worldwide), engineers must adopt a systems thinking approach to innovation that allows them to understand the sociotechnicity of AI systems across sociocultural domains, and that (2) the value sensitive design (VSD) approach, and in particular *envisioning cards*, provides a decisive first step towards helping designers harmonise human values, understood across geospatial and temporal boundaries, with economic values, rather than the former coming at the opportunity cost of the latter.

The paper is thus organised into the following sections. §2 outlines systems theory and systems engineering, as well as how this way of thinking provides a more accurate ontological understanding of designing *for* human values, rather than relegating these values to the position of mere afterthoughts. §3 outlines the VSD approach as a method of systems engineering that guides designers in harmonising human values during the design process. §4 discusses the VSD technique known as *envisioning cards* in greater depth and argues that they provide a solid initial step towards AI design that harmonises the long-term consideration of human values with more immediate economic values. §5 discusses the limitations of this approach and highlights some areas for future research projects. Finally, §6 concludes this paper with a summary of its contribution.

2. Systems Thinking And Systems Engineering

Although the meaning of the term 'systems theory' seems self-evident, it warrants closer analysis. To draw on systems theory as a conceptual framework fundamentally means making ontological commitments to understanding AI. Although I will not include a full discussion of systems theory and its long history here, there are numerous reasons for adopting it. Firstly, systems theory characterises the nuances of complex systems that describe our increasingly complex sociotechnical worlds. AI systems further complicate these matters and their analysis requires a conceptual language that can help us navigate these complexities. Secondly, systems engineering is the practical endeavour of designing systems and is built upon the theoretical underpinnings of systems theory. Finally, as discussed later on in this paper, VSD is essentially a systems engineering approach to technological design. As such, a discussion of ontological substrata will equip us with a better understanding of how salient AI design can take place. Finally, because it affirms an *interactional* stance on technology, VSD acknowledges that technology and societal forces co-construct and co-vary with one another (Friedman et al., 2017). This means that purely deterministic, instrumental, or constructivist understandings of technologies are not entirely correct when taken individually. Instead, the plurality of actors, institutions, and technologies, and their design histories, form complex yet important networks of interactions. These relationships need to be brought to the fore in order to create the conditions for responsible innovation.

2.1 Systems Theory

A more widely conceptualised field, 'systems theory' has been defined as the interdisciplinary study of organised and complex systems (Whitchurch & Constantine, 2009). Although systems often draw from specimens in the biological domain, they can also be synthetic and understood as connected clusters of co-constitutive and co-varying nodes. Spatiotemporal vectors bind these clusters together, while their environment transforms them. They are defined by their composition and teleology, the latter of which is expressed through operation (Adams et al., 2014). Thinking along these lines is described as 'systems thinking', which is simply the verbal conceptualisation of things as functions of systems theory axioms. Taking this on board, the theory often describes systems very differently in comparison to other complex relational ontologies, like object-oriented ontology (OOO). Emergent behaviours (or synergy) often leads to the description of systems as being *more* than the sum of their parts, whereas theories like OOO lead to the opposite conclusion (Haken, 2013; Harman, 2018). Furthermore, due to the complexity of systems, modulations at any vector or node within the system can generate cascading changes throughout the system at other vectors or nodes. This can result in unforeseen or even unforeseeable emergent behaviour (akin to what we already see in various opaque AI systems) (Floridi et al., 2020; Hibbard, 2012). Hence, systems theory studies the patterns of connections and complexity to predict future behaviour more accurately.

In general, given the number of inputs, nodes, and potential emergent behaviours, sociotechnical systems are rendered even more complex by self-learning and adaptive systems, such as those that characterise AI systems. Because the term 'AI' is often used haphazardly and to refer to systems that are not relevant here, I adopt the definition of 'AI' used by Umbrello and van de Poel (2021) as the "class of technologies that are autonomous, interactive, adaptive, and capable of carrying out human-like tasks (Floridi & Sanders, 2004) […] particularly […] AI technologies based on Machine Learning (ML), which allows such technologies to learn based on interaction with (and feedback from) the environment" (Umbrello & van de Poel, 2021, p.1). This interplay between the system and its environment also means that *systems are within systems* and that such interplay supports or constrains certain behaviours, thus making the systems more or less robust. Part of systems thinking, then, is understanding the kinetics, interplay, and bounding conditions. This allows more pertinent extrapolations to be made to improve how we conceptualise other systems across levels of recursion (Graham et al., 1994).

General systems theory (GST) aims to synthesise methods and instruments to create a broader understanding of complex systems, as opposed to siloed disciplinary approaches (von Bertalanffy, 1972). GST further classifies systems into two categories: active and passive. The former is understood as a system that engages in processes and exhibits dynamic behaviour, whereas the latter is engaged by active systems. If we take, for example, an AI-powered auditing system, it is passive when it is neither activated nor processing information, whereas when booted, it becomes an active system. Hence, the spatiotemporal vector mentioned above is essential because it determines whether we describe a system as passive or active. Importantly, the nodes that make up a system or that operate within a more extensive system can likewise be characterised as active or passive components. This is significant because, when considering the complexity of AI systems, particularly those that employ machine learning (ML) (or ML and artificial neural network hybrids), the algorithmic process is often opaque (Turilli & Floridi, 2009). To account for this complexity and propose relevant systems design, we need a GST approach to map the sophisticated connections and nuances that characterise AI systems.

2.2 Systems Engineering.

Systems engineering can be understood as applying a systems thinking (and thus transdisciplinary) approach to engineered systems. It uses systems thinking to understand, design, manage, and deploy engineered systems to ensure optimised equifinality over their lifecycles (Thomé, 1993). This approach, however, is not purely technical, but also incorporates humanistic disciplines such as organisational studies, ethics, and project management (Booton & Ramo, 1984). A more holistic, comprehensive approach to systems design is therefore possible when we frame systems not as extricated artefacts but as part of situated environments in which they play an integral role. Furthermore, merging multiple disciplines enables greater synergism between constituent parts, helping designers predict future emergent behaviours more precisely (Bauer & Herder, 2009).

As it features a more comprehensive understanding of *systems-within-systems*, systems engineering is therefore orientated towards optimising equifinality because it views complex technologies as dynamic systems with emergent behaviours. This means that systems cannot simply be designed and deployed without further consideration; instead, they require co-design and monitoring over their entire lifecycle (SyntheSys, 2020; Umbrello & van de Poel, 2021).

As new emergent behaviours become manifest over time, so too will new design requirements develop in response to changing values (van de Poel, 2018, 2020). As we shall see with VSD, systems engineering models confirm that system behaviours result from their architecture. When assembled and organised in a particular environment, individual nodes constitute the 'black box' of elements that define the system within a system. The complexity of these systems and their interplay can be expounded to multiple levels of abstraction. An example of this is a medical diagnostic system that is composed of various systems and forms part of numerous interconnected systems (i.e., a hospital information and communications technology network). By modelling systems, designers can engineer them to have more predictable (desirable) behaviours in their intended environment. Likewise, mapping system processes and behaviours also enables unforeseen behaviours to be addressed and ameliorated early on in the design process and throughout. This kind of engineering approach to systems enables consideration of the most salient aspects of systems verification, integration, and validation, rather than waiting for recalcitrant behaviour to occur *in situ* and incur unwanted costs (both social and economic).

3. Vsd, Interactionalism And Systems Engineering

As a method of technology design, VSD is often described as a 'principled approach', given its overt orientation towards designing technologies *for* human values rather than consigning them to *ad hoc* afterthoughts (Friedman, 1996). With almost 30 years of history and development underlying the approach, co-creation between both direct and indirect stakeholders[1] is a fundamental part of the design process, as is the philosophical investigation of values (Umbrello, 2018). Past research has explored how VSD can be applied to specific technologies, such as energy transition systems (Mok & Hyysalo, 2018), mobile phone usage (Woelfer et al., 2011), industrial processes (Longo et al., 2020), and more recent systems of augmented reality (Friedman & Kahn Jr., 2000), to name just a few. It has similarly been proposed as a suitable design framework for future technologies, both in the short and long term. Examples include its exploratory application to nanopharmaceuticals (Timmermans et al., 2011), molecular manufacturing (Umbrello, 2019), care robots (Umbrello et al., 2021; van Wynsberghe, 2013), and less futuristic autonomous vehicles (Calvert et al., 2018; Thornton et al., 2018; Umbrello & Yampolskiy, 2021).

Despite all these uses, VSD has only been applied to AI systems *conceptually*, as AI's self-learning capabilities pose some unique challenges for the VSD approach. To combat these, Umbrello and van de Poel (2021) suggest adding a set of AI-specific design principles to VSD predicated on the advancements made in the various AI for Social Good (AI4SG) projects (Mabaso, 2020; Taddeo & Floridi, 2018). However, even these more specific norms are insufficient and require additional value sources that can be harmonised with the intention of designing AI4SG using VSD. Stakeholder values represent one such source, which are constituent of 'context analysis' in the authors' four-stage VSD approach. They argue that context is crucial in all AI design:

In all cases […], different contextual variables come into play to impact the way values are understood (in the second phase), both in conceptual terms as well as in practice, on account of different sociocultural and political norms. The VSD approach sees eliciting stakeholders in sociocultural contexts as imperative. This will determine whether the explicated values of the project are faithful to those of the stakeholders, both directly and indirectly. Empirical investigations thus play a key role in determining potential boons and downfalls for any given context. (Umbrello & van de Poel, 2021, p. 7).

To understand the importance of this, both to VSD more broadly and the design of the AI system in particular, the inner workings of VSD merit brief discussion. Sometimes heralded under the auspices of somewhat different names, such as 'Values at Play' or 'Design for Values' (Flanagan & Nissenbaum, 2014; van den Hoven et al., 2015), VSD is traditionally described as a three-phase methodology comprising conceptual, empirical, and technical investigations (van den Hoven & Manders-Huits, 2009). Moreover, the tripartite approach can be engaged with iteratively or consecutively (see Fig. 1).

Conceptual investigations involve *a priori* analysis of the potential value implications and identification of direct and indirect stakeholders, as well as the likely value tensions. This phase also involves coming up with working definitions of values that can then inform (and be informed by) the other investigations. Empirical investigations involve eliciting data from the stakeholders themselves in an attempt to determine their values and value understandings. This information feeds back into the other phases to help refine the working definition of the 'value at play'. Finally, technical investigations look at the technology itself, or, more specifically, how the architecture and design choices of the system might support and/or constrain those values.

Philosophically speaking, the entire VSD approach is premised on the *interactional* stance regarding technology. VSD thus argues against the value-neutrality thesis of technology (i.e., *instrumentalism*) and instead claims that technologies embody the values of their creators. This means that they display properties that are both deterministic as well as constructionist (Friedman & Hendry, 2019). This is a salient way of understanding technological artefacts' sociotechnicity (as in the case of Winner's bridges). Societal forces and technologies co-construct, co-vary, and co-constitute each other (Ropohl, 1999). VSD is currently equipped with seventeen specific methods to facilitate systems design in light of sociotechnicity: (1) stakeholder analysis; (2) stakeholder tokens; (3) value source analysis; (4) coevolution of technology and social structure; (5) value scenarios; (6) value sketches; (7) value-oriented semi-structured interview; (8) scalable assessments of information dimensions; (9) value-oriented coding manual; (10) value-oriented mock-ups, prototypes, and field deployments; (11) ethnography focused on values and technology; (12) model for informed consent online; (13) value dams and flows; (14) value sensitive action-reflection model; (15) multi-lifespan timeline; (16) multi-lifespan co-design; and (17) envisioning cards (Friedman & Hendry, 2019).

To achieve the objective of designing *for* human values, these methods each have their own uses. These include stakeholder identification and legitimation, value source identification and definition, determining how such values relate to their contextual social structures, and design thinking across multiple generations. The suitability of any one method is contingent on the starting point of any given engineering programme. However, part of the attractiveness of VSD is that it can and should be adapted to an individual domain of application. Crucially, it is not a wholesale reimagining of the design space, but instead maps onto and augments existing design and engineering practices. This is an important point: AI systems design is advancing at a remarkable pace globally, and because firms recognise the economic and other market advantages of adopting AI systems, they are more than willing to adopt less-than-ready systems *despite* the potential for recalcitrance (see e.g., Banerjee & Chanda, 2020). As a result, an adaptable design approach that can be cost-effectively mapped onto existing design practices is invaluable. Although little work has been done on this point regarding VSD, a clear objective of this design methodology is that it should not replace but rather complement the day-to-day practices of technology designers (Friedman & Hendry, 2019; van de Poel, 2018). Although specific VSD tools may indeed take more time to implement than others, there are nonetheless VSD methods available to AI systems designers that can help them avoid many pitfalls caused by short-term, market imperative thinking.

I am here referring primarily to VSD's systems-oriented approach. An explicit aspect of the interactional stance on technology means looking not only at discrete technologies but also viewing them as fundamental and inseparable constituents of social forces, organisations, institutions, and infrastructures (i.e., as *systems-within-systems*). Likewise, VSD takes a complete systems view of this broader design context by including the various direct and indirect stakeholders implicated in these systems. As mentioned above, VSD, like systems engineering generally, draws on the theories and methods of multiple disciplines to achieve greater equifinality in design. Mapping out the long-term network effects that the system can produce is therefore necessary, though it runs contrary to much of the previously discussed short-term thinking that characterises most modern innovation practices. This short termism cannot be risked with transformative technologies like AI.

The following section proposes the use of *envisioning cards* as an easily adoptable way for AI design firms to engage in VSD while also minimising drastic internal changes, thus harmonising their economic incentives with critical human values. This approach permits long-term, multi-generational thinking for a wider group of stakeholders, which is highly relevant for AI and other globally impactful sociotechnical systems.

[1] 'Direct stakeholders' are those who may be impacted via direct interaction with the technology. They can include users, designers, and some managers. 'Indirect stakeholders' are those who may be impacted by the systems but do not directly interact with it. They can include stakeholder groups like executives, other publics, and the environment or nonhuman animals.

4. Envisioning Vsd For Ai Systems Design

AI systems, particularly those based on ML, are markedly sociotechnical systems. This means that they are inextricably part of their social contexts. They also form part of larger systems, including cyberspace and information and communications technology, and are thus also *systems-within-systems* (Umbrello & Gambelin, 2022; Uphoff, 2014). Beyond this, and unlike other, more discrete sociotechnical systems with a more stable spatiotemporal locus, AI is by nature easily diffused globally and implicates large stakeholder populations that were unaccounted for early in the design of these systems (Floridi et al., 2020). Likewise, the unique challenges posed to salient design by opaque processes make mapping emergent behaviours difficult, though necessary.

So far, this paper has outlined systems theory and engineering, showing how they address both the ontology and ethics of designing AI systems. The former describes the systemic nature, connectivity, and emergence that characterise AI systems, while the latter explains *how* to design for complexity and equifinality. I argue that VSD is one such approach to systems engineering, uniquely capable of designing within the systems thinking paradigm (c.f., Umbrello et al., 2021). It is also apt at meeting the exceptional challenges posed by AI systems, while simultaneously thinking long term and across spatiotemporal boundaries to include wider stakeholder communities. As a strong starting point, one VSD method that AI design firms can adopt at a relatively low cost, in terms of both time and money, is *envisioning cards* (Friedman et al., 2017).

4.1 The Envisioning Criteria

Built on more than two decades of conceptual and empirical work within VSD, *envisioning cards* represent one of seventeen existent VSD methods that can be adopted with the goal of designing systems *for* values (Friedman et al., 2017). Although VSD is more generally a long-term approach to design, *envisioning cards* are intended to stress the unique challenges of long-term thinking and provide actionable means to address those challenges. Each set of *envisioning cards* comes with two decks: (1) the primary set (28 cards + 4 'create your own' cards) and (2) the supplementary multi-lifespan set (12 cards + 1 ‘create your own’ card). All the cards are adorned with a provocative image on one side, with the other side displaying the *envisioning criterion*, *title*, *description/theme*, and a *design activity* or actionable *prompt*. The primary set of cards includes four *envisioning criteria*; each card highlights one of those criteria, provides a thoughtful description of the issues associated with that criterion, and prompts the user to consider how to tackle potential issues with a creative prompt or design activity (see Fig. 2).

Each of the primary deck’s four criteria highlights different aspects of issues that may emerge due to the design choices in any given system. Figure 3 illustrates each envisioning criterion.

However, the supplementary set has an extra criterion in addition to the four existing ones: *multi-lifespan* (Fig. 4).

The additional set draws on over a decade’s worth of VSD research on the design of information systems across multiple lifespans (Friedman et al., 2016; Friedman & Yoo, 2017; Yoo et al., 2016). AI systems are already globally widespread, and their ubiquitous uptake and consequent sociotechnical pervasiveness undoubtedly mean that they will continue to exist for many generations to come. Any relevant attempt to address these considerable structural challenges thus requires similarly extended timeframes. The *multi-lifespan envisioning cards* are primarily orientated towards designing systems that pose these significant societal issues but resist expedient fixes. This supplementary set of cards encourages designers to consider design choices in the present day that are consciously and explicitly directed towards policies, infrastructures, and systems architectures that would open up the most comprehensive array of design options for future generations (c.f., van den Hoven, 2017).

Individual designers or design teams can use the *envisioning cards* for various ends, such as finding creative solutions to potentially intractable problems, determining novel criteria for success in a design, assessing the value tensions of clients, and widening the scope of potentially impacted stakeholder populations. Like VSD more broadly, the *envisioning cards* are not intended as a wholesale reimagining of the design domain in which they are used. If they were, they would present, as with any potential approach, a high barrier to entry, thus negatively impacting their adoption and therefore the potential value derived from their use. Instead, the *envisioning cards* are meant to seamlessly map onto existing design practices regardless of the approach or process being adapted. For example, many software development firms employ some form of Agile or Waterfall workflow management for their design projects. VSD in general, and *envisioning cards* more specifically, can act as a vehicle for values without burdening firms with further financial or time constraints, which may result from other techniques used to retool their normal day-to-day activities. For example, Umbrello and Gambelin (2022) argue that the VSD approach and its various methods can be easily understood as elements of existing Agile phases (Fig. 5). A similar process for reframing VSD as a tool for these existing workflows satisfies VSD’s internal philosophical precepts of seamless applicability as well as resistance to short-termism, the latter of which is characteristic of methodologies like Agile and Waterfall (Umbrello & Gambelin, 2022).

This point is significant: much of the AI development domain fits squarely within corporate structures characterised by their use of short-term project management approaches whose success is often measured in terms of return on investment. Often this comes at the opportunity cost of the value-centred design of AI products (and systems in general). This is primarily the consequence of trade-off thinking, something that VSD is philosophically predicated against. VSD is built on the notion that most innovations are developed within a sphere where economic values are front and centre. However, VSD does not argue that moral values come at the opportunity cost of economic ones; in fact, the opposite is true. VSD claims that not only do they complement each other, but they also augment each other as a consequence of creative design. For example, Sweden's zero-tolerance policy for road accidents has led to innovative safety technologies being implemented to meet these strict requirements (Kristianssen et al., 2018). As a result, automotive manufacturers like Volvo have become leaders in automotive safety. Rather than positioning safety at the opportunity cost of economic profit, it is framed as a necessary prerequisite for economic value. Greater safety leads to bigger profit.

More generally, the *envisioning cards* provide firms with an easy-to-adopt approach at a marginal cost, which is capable of being used in the design of AI systems that encompass global stakeholder groups across spatiotemporal vectors (across the world and across lifespans). Many of the activities take less than three minutes to complete. For example, Figure 6, the 'Remembering and Forgetting' *multi-lifespan envisioning card*, can be quickly geared towards long-term thinking in AI design.

'Remembering and Forgetting' is particularly salient in the context of AI design. It can be framed as highlighting issues with data storage and recall in AI systems, particularly the data sets that are being used to train and run more significant ML and artificial neural network (ANN)-based systems (e.g., Stoica et al., 2017). The theme of the card describes issues of data regulation, access (and by whom), types of data access, as well as data destruction (i.e., forgetting, the right to be forgotten) (see e.g., Rosen, 2011). The card prompt follows suit with a direction actionable exercise to stimulate long-term thinking regarding the impacts of this kind of information storage/use, in this case up to 50 years into the future. Other *multi-lifespan envisioning cards* like 'Cultivating Trust' and 'Material Longevity' (see Fig. 7) also fit nicely within the current discussions on issues arising from the design and deployment of AI systems.

For instance, 'Cultivating Trust' helps designers imagine how the use of the system may compromise stakeholders and asks how trust can be strengthened over time. 'Material Longevity' allows designers to determine long-term viability, given the current materials necessary for the system's operation, by encouraging them to list materials the system relies on and determine the characteristics and impacts of the design choice of those materials.

5 Limitations And Avenues For Future Research

Overall, this paper has addressed some of the fundamental issues regarding AI design. Firstly, the design and development of AI systems is primarily undertaken by private firms where economic values are more often than not framed as being prioritised at the opportunity cost of morally important human values like *human autonomy*, *fairness, nonmaleficence*, and *explicability,* among many others. Similarly, by discussing systems theory/thinking, we can see how the sociotechnicity of AI systems poses some unique challenges, particularly when we consider the long-term impacts of today's design choices. *Envisioning cards* represent one of seventeen existing VSD tools that AI design firms can adopt to begin designing *for* human values rather than waiting for the emergent and potentially recalcitrant behaviour of these systems to appear. This paper proposes *envisioning cards* as a potentially fruitful starting point, given that they are relatively low cost and provide easy-to-use activities that can be integrated into any existing project management process.

This, however, does not entail that *envisioning cards* alone are sufficient to produce or ensure value-aligned AI systems across time. At the very least, they help bring to the surface unforeseen values, stakeholders, and multi-generational impacts that these systems may manifest if deployed. It is up to designers and engineers to then determine the technical means by which such systems can address or ameliorate these issues before they develop. Other approaches like *value scenarios* may be a practical next step for firms interested in more granular and targeted AI design (Nathan et al., 2007). Similarly, a multi-tiered VSD approach like that proposed by Umbrello and van de Poel (2021) may be a further step. However, these more intensive approaches come with the cost of specialised training and expertise, meaning they do not necessarily provide the ideal entry-level step for firms who may already be hesitant to change their current practices.

6. Conclusions

AI systems are autonomous, interactive, adaptive, and capable of carrying out human-like tasks. These types of systems are already witnessing ubiquitous uptake across the globe and they are generating various impacts worldwide. As this uptake becomes more pervasive, so will the systemic effects of their emergent behaviours. This poses some unique ethical issues for design. Many of the loci of AI innovation are spatiotemporally situated in individual development firms with narrow considerations for the impacted stakeholders. In this paper, I have described AI as a paragon of the systems thinking approach to technology. In doing so, I argue that the sociotechnicity of systems requires a design approach that is fundamentally congruent with this systems ontology, that is, systems engineering. I propose VSD as one such systems engineering approach that addresses the singular challenges posed by AI design. Finally, *envisioning cards*, one of the VSD methodologies, is suggested as an easily adoptable first step for AI design firms to begin setting the stage for value-sensitive AI design and deployment.

Declarations

Funding: this research received no specific funding.

Conflicts of interest: The authors declare no conflict of interest

Authors' contributions: All authors contributed equally to the organization, drafting and editing of this article.

Acknowledgments: The views expressed in this article do not reflect those of the Institute for Ethics and Emerging Technologies.

References

1. Adams, K. M., Hester, P. T., Bradley, J. M., Meyers, T. J., & Keating, C. B. (2014). Systems theory as the foundation for understanding systems. *Systems Engineering*, *17*(1), 112–123.
2. Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines*, *22*(4), 299–324. https://doi.org/10.1007/s11023-012-9282-2
3. Banerjee, D. N., & Chanda, S. S. (2020). *AI failures: A review of underlying issues*. http://arxiv.org/abs/2008.04073
4. Bauer, J. M., & Herder, P. M. (2009). Designing socio-technical systems. In *Philosophy of technology and engineering sciences*, Anthonie Meijers (ed.) (pp. 601–630). North Holland: Elsevier.
5. Booton, R. C., & Ramo, S. (1984). The development of systems engineering. *IEEE Transactions on Aerospace and Electronic Systems*, *4*, 306–310.
6. Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, *22*(2), 71–85. https://doi.org/10.1007/s11023-012-9281-3
7. Calvert, S. C., Mecacci, G., Heikoop, D. D., & de Sio, F. S. (2018). Full platoon control in truck platooning: A meaningful human control perspective. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 3320–3326.
8. Flanagan, M., & Nissenbaum, H. (2014). *Values at play in digital games*. MIT Press.
9. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5
10. Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics*, *26*(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5
11. Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349–379.
12. Friedman, B. (1996). Value-sensitive design. *Interactions*, *3*(6), 16–23. https://doi.org/10.1145/242485.242493
13. Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press.
14. Friedman, B., Hendry, D. G., & Borning, A. (2017). A survey of value sensitive design methods. *Foundations and Trends® in Human–Computer Interaction*, *11*(2), 63–125. https://doi.org/10.1561/1100000015

15. Friedman, B., & Kahn Jr., P. H. (2000). New directions: A value-sensitive design approach to augmented reality. *Proceedings of DARE 2000 on Designing Augmented Reality Environments*, 163–164. https://doi.org/10.1145/354666.354694
16. Friedman, B., Nathan, L., Kane, S., & Lin, J. (n.d.). *The envisioning criteria*. Envisioning Cards. Retrieved June 16, 2020, from https://www.envisioningcards.com/?page_id=2#5
17. Friedman, B., Nathan, L. P., Kane, S. K., & Lin, J. (2017). *Envisioning cards*. University of Washington, Value Sensitive Design Research Lab. https://www.envisioningcards.com
18. Friedman, B., Nathan, L. P., & Yoo, D. (2016). Multi-lifespan information system design in support of transitional justice: Evolving situated design principles for the long (er) term. *Interacting with Computers*, *29*(1), 80–96.
19. Friedman, B., & Yoo, D. (2017). Pause: A multi-lifespan design mechanism. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 460–464.
20. Graham, R., Knuth, D., & Patashnik, O. (1994). 1. Recurrent Problems. In *Concrete Mathematics: A Foundation for Computer Science*, Graham, R. L., Knuth, D. E., Patashnik, O. (Eds.) (2nd ed., p. 670). Reading, MA: Addison-Wesley Professional.
21. Haken, H. (2013). *Synergetics: Introduction and advanced topics*. Springer Science & Business Media.
22. Harman, G. (2018). *Object-oriented ontology: A new theory of everything*. Penguin Random House.
23. Hibbard, B. (2012). Avoiding unintended AI behaviors. In J. Bach, B. Goertzel, & I. Matthew (Eds.), *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): Vol. 7716 LNAI* (pp. 107–116). Springer. https://doi.org/10.1007/978-3-642-35506-6_12
24. Kristianssen, A. C., Andersson, R., Belin, M. Å., & Nilsen, P. (2018). Swedish Vision Zero policies for safety – A comparative policy content analysis. *Safety Science*, *103*, 260–269. https://doi.org/10.1016/j.ssci.2017.11.005
25. Longo, F., Padovano, A., & Umbrello, S. (2020). Value-oriented and ethical technology engineering in industry 5.0: A human-centric perspective for the design of the factory of the future. *Applied Sciences (Switzerland)*, *10*(12), 1–25. https://doi.org/10.3390/APP10124182
26. Mabaso, B. A. (2020). Artificial moral agents within an ethos of AI4SG. *Philosophy & Technology*, 1–15.
27. Mok, L., & Hyysalo, S. (2018). Designing for energy transition through value sensitive design. *Design Studies*, *54*, 162–183.
28. Nathan, L. P., Klasnja, P. V, & Friedman, B. (2007). Value scenarios: A technique for envisioning systemic effects of new technologies. *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, 2585–2590. https://doi.org/10.1145/1240866.1241046
29. Ropohl, G. (1999). Philosophy of socio-technical systems. *Techne: Research in Philosophy and Technology*, *4*(3), 186–194. https://doi.org/10.5840/techne19994311

30. Rosen, J. (2011). The right to be forgotten. *Stanford Law Review Online*, *64*. https://heinonline.org/HOL/Page?handle=hein.journals/slro64&id=89&div=17&collection=journals
31. Stahl, B. C. (2004). Information, ethics, and computers: The problem of autonomous moral agents. *Minds and Machines*, *14*(1), 67–83. https://doi.org/10.1023/B:MIND.0000005136.61217.93
32. Stoica, I., Song, D., Popa, R. A., Patterson, D., Mahoney, M. W., Katz, R., Joseph, A. D., Jordan, M., Hellerstein, J. M., Gonzalez, J. E., Goldberg, K., Ghodsi, A., Culler, D., & Abbeel, P. (2017). *A Berkeley view of systems challenges for AI*. http://arxiv.org/abs/1712.05855
33. SyntheSys. (2020, July). Why use systems engineering? *The IT Insider*. https://theitinsider.co.uk/articles/2020/why-use-systems-engineering/
34. Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*(6404), 751–752.
35. Thomé, B. (1993). *Systems engineering: principles and practice of computer-based systems engineering*. John Wiley and Sons.
36. Thornton, S. M., Lewis, F. E., Zhang, V., Kochenderfer, M. J., & Gerdes, J. C. (2018). Value sensitive design for autonomous vehicle motion planning. *2018 IEEE Intelligent Vehicles Symposium (IV)*, 1157–1162.
37. Timmermans, J., Zhao, Y., & van den Hoven, J. (2011). Ethics and nanopharmacy: Value sensitive design of new drugs. *NanoEthics*, *5*(3), 269–283. https://doi.org/10.1007/s11569-011-0135-x
38. Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*(2), 105–112. https://doi.org/10.1007/s10676-009-9187-9
39. Umbrello, S. (2018). The moral psychology of value sensitive design: The methodological issues of moral intuitions for responsible innovation. *Journal of Responsible Innovation*, *5*(2), 186–200. https://doi.org/10.1080/23299460.2018.1457401
40. Umbrello, S. (2019). Atomically precise manufacturing and responsible innovation: A value sensitive design approach to explorative nanophilosophy. *International Journal of Technoethics*, *10*(2), 1–21. https://doi.org/10.4018/IJT.2019070101
41. Umbrello, S. (2020). Meaningful human control over smart home systems: A value sensitive design approach. *Humana.Mente Journal of Philosophical Studies*, *13*(37), 40–65.
42. Umbrello, S., Capasso, M., Balistreri, M., Pirni, A., & Merenda, F. (2021). Value sensitive design to achieve the UN SDGs with AI: A case of elderly care robots. *Minds and Machines*, 1–25. https://doi.org/10.1007/s11023-021-09561-y
43. Umbrello, S., & Gambelin, O. (2022). *Agile as a vehicle for values: A value sensitive design toolkit. Working paper.*
44. Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00038-3
45. Umbrello, S., & Yampolskiy, R. V. (2021). Designing AI for explainability and verifiability: A value sensitive design approach to avoid artificial stupidity in autonomous vehicles. *International Journal of Social Robotics*. https://doi.org/10.13140/RG.2.2.10855.68003

46. Uphoff, N. (2014). Systems thinking on intensification and sustainability: Systems boundaries, processes and dimensions. *Current opinion in environmental sustainability 8*, 89–100. https://doi.org/10.1016/j.cosust.2014.10.010
47. van de Poel, I. (2018). Design for value change. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-018-9461-9
48. van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, *30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4
49. van den Hoven, J. (2017). The design turn in applied ethics. In J. van den Hoven, S. Miller, & T. Pogge (Eds.), *Designing in ethics* (pp. 11–31). Cambridge University Press. https://doi.org/10.1017/9780511844317
50. van den Hoven, J., & Manders-Huits, N. (2009). Value-sensitive design. In J. K. B. Olsen, S. A. Pedersen, & V. F. Hendricks (Eds.), *A companion to the philosophy of technology*. Wiley-Blackwell, 477-480. https://doi.org/10.1002/9781444310795.ch86
51. van den Hoven, J., Vermaas, P. E., & van de Poel, I. (Eds.) (2015). *Handbook of ethics, values, and technological design: Sources, theory, values and application domains*. Springer Netherlands. https://doi.org/10.1007/978-94-007-6970-0
52. van Wynsberghe, A. (2013). Designing robots for care: Care centered value-sensitive design. *Science and Engineering Ethics*, *19*(2), 407–433. https://doi.org/10.1007/s11948-011-9343-6
53. Verbeek, P.-P. (2008). Disclosing visions of technology. *Techné: Research in Philosophy and Technology*, *12*(1), 85–89. https://doi.org/10.5840/techne200812116
54. von Bertalanffy, L. (1972). The history and status of general systems theory. *Academy of Management Journal*, *15*(4), 407–426.
55. Whitchurch, G. G., & Constantine, L. L. (2009). Systems theory. In *Sourcebook of family theories and methods*, Boss, P., Doherty, W.J., LaRossa, R., Schumm, W.R., Steinmetz, S.K. (Eds.) (pp. 325–355). New York: Springer.
56. Winner, L. (2003). Do artifacts have politics? *Technology and the Future*, *109*(1), 148–164. https://doi.org/10.2307/20024652
57. Woelfer, J. P., Iverson, A., Hendry, D. G., Friedman, B., & Gill, B. T. (2011). Improving the safety of homeless young people with mobile phones: Values, form and function. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1707–1716. https://doi.org/10.1145/1978942.1979191
58. Yoo, D., Logler, N., Ballard, S., and Friedman, B. (2018). *Multi-lifespan envisioning cards – supplementary set.* Value Sensitive Design Lab, University of Washington, Seattle, WA. Available at: https://www.envisioningcards.com/
59. Yoo, D., Derthick, K., Ghassemian, S., Hakizimana, J., Gill, B., & Friedman, B. (2016). Multi-lifespan design thinking: Two methods and a case study with the Rwandan diaspora. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4423–4434.

Figures

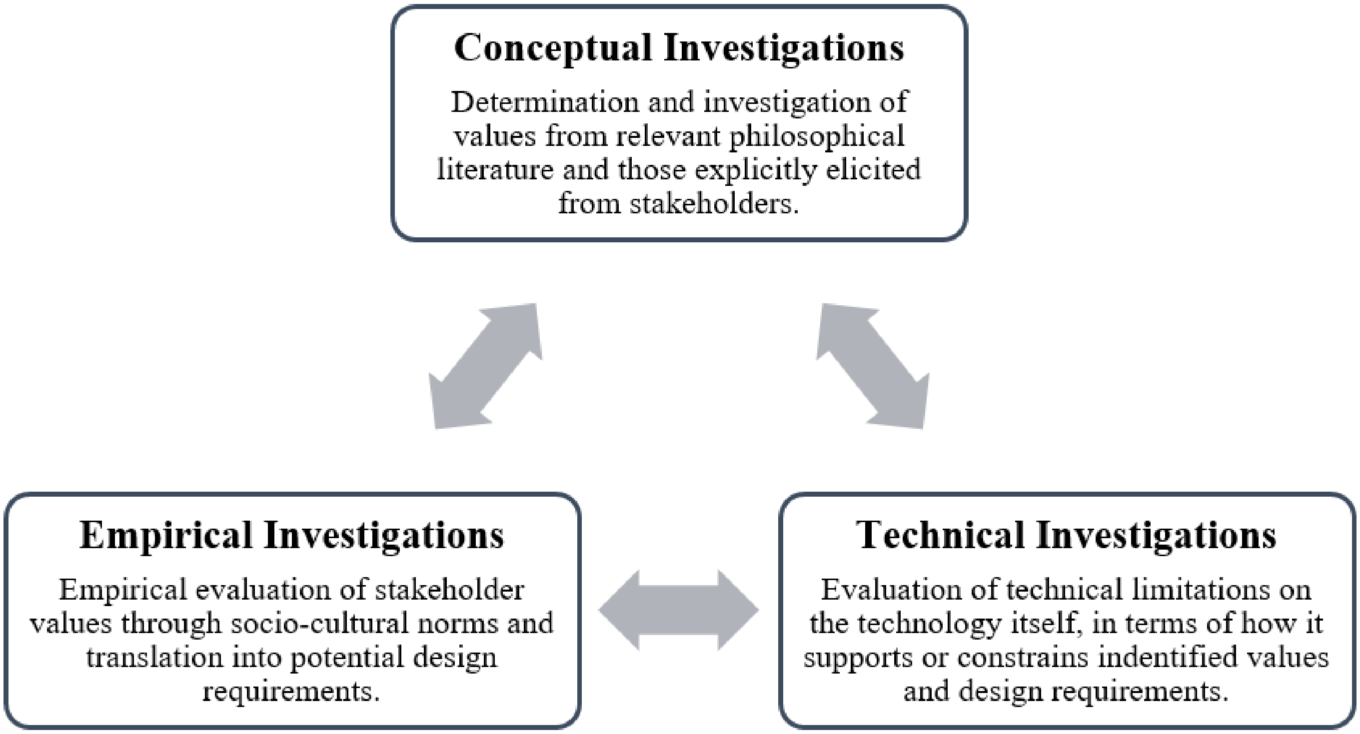

Figure 1

The recursive VSD tripartite framework employed in this study. Source Umbrello (2020)

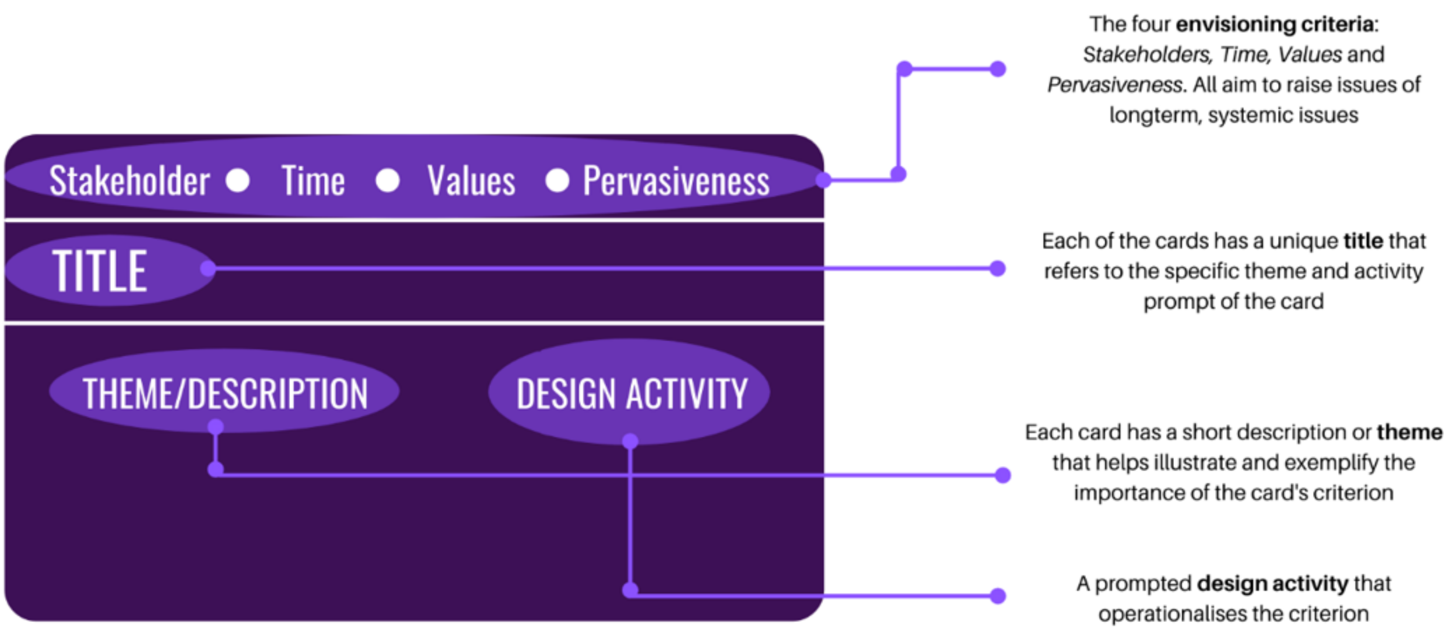

Figure 2

Description of contents of the primary set of envisioning cards.

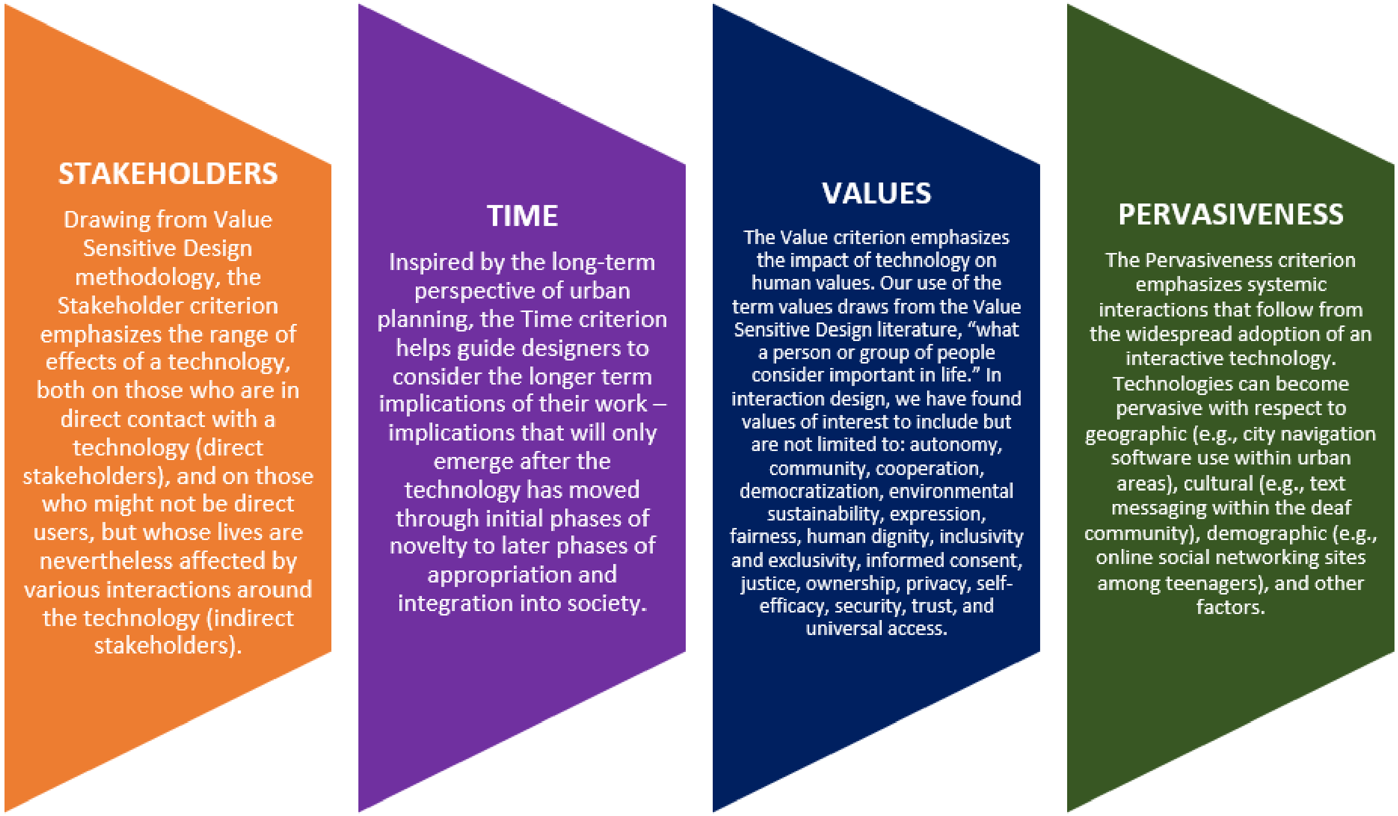

Figure 3

The envisioning criteria. Source Friedman et al. (n.d.)

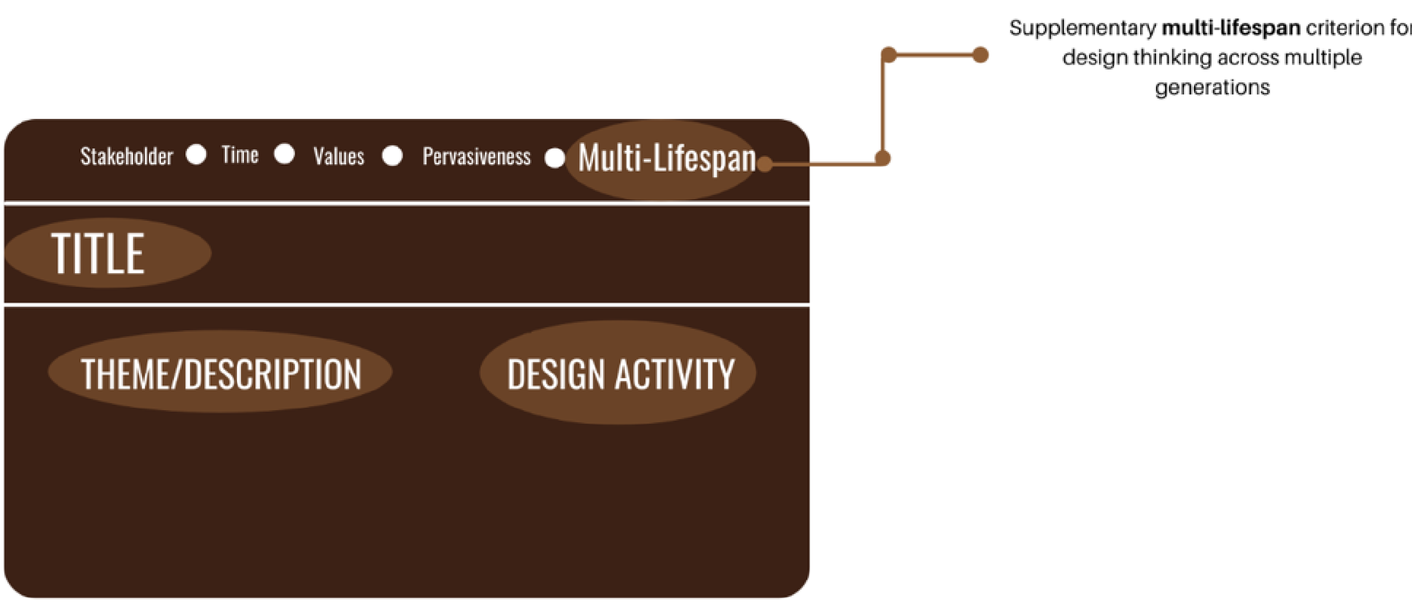

Figure 4

Description of supplementary multi-lifespan envisioning cards set.

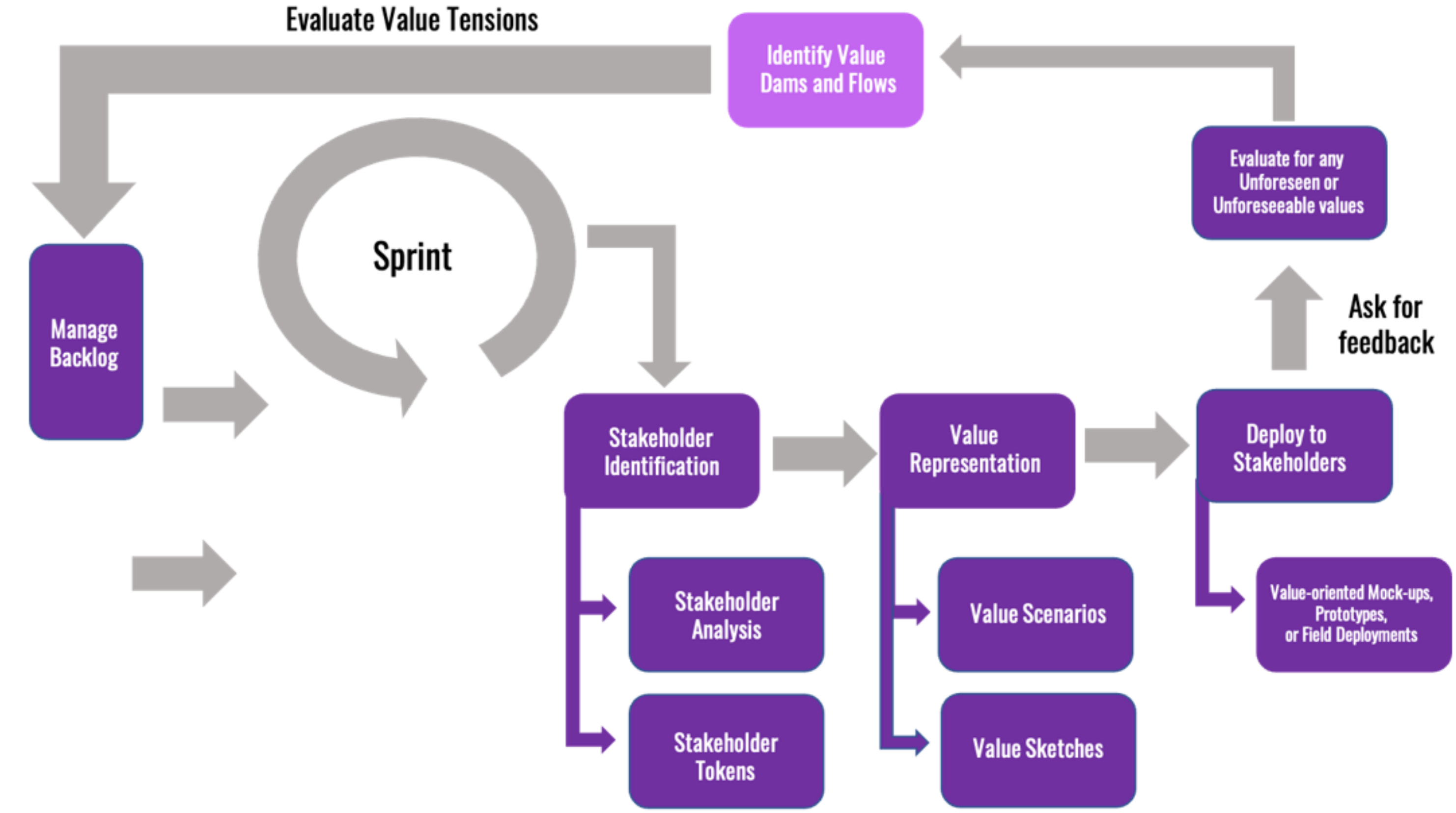

Figure 5

Agile workflow with the addition of VSD value representation tools. Source Umbrello & Gambelin (2022)

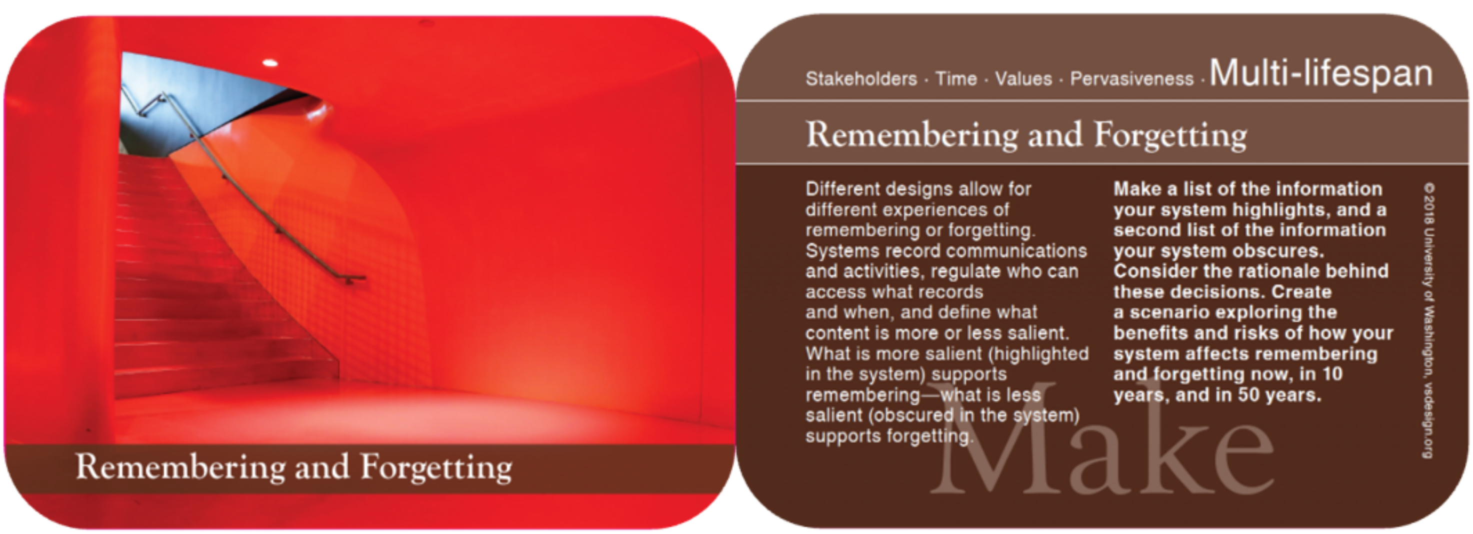

Figure 6

'Remembering and Forgetting' multi-lifespan envisioning card. Source Yoo et al. (2008)

Figure 7

'Cultivating Trust' and 'Material Longevity' multi-lifespan envisioning card. Source Yoo et al. (2018)