Conceptualizing Policy in Value Sensitive Design:

A Machine Ethics Approach

Steven Umbrello

*Institute for Ethics and Emerging Technologies; University of Turin, Italy*

*steven.umbrello@unito.it*

**Abstract**

*The value sensitive design (VSD) approach to designing emerging technologies for human values is taken as the object of study in this chapter. VSD has traditionally been conceptualized as another type of technology or instrumentally as a tool. The various parts of VSD’s principled approach would then aim to discern the various policy requirements that any given technological artifact under consideration would implicate. Yet, little to no consideration has been given to how laws, policies and social norms engage within VSD practices. Similarly, how the interactive nature of the VSD approach can, in turn, influence those directives. This is exacerbated when considering machine ethics policy that has global consequences outside their development spheres. This chapter begins with the VSD approach and aims to determine how policies come to influence how values can be managed within VSD practices. It shows that the interactional nature of VSD permits and encourages existing policies to be integrated early on and throughout the design process.*

Keywords: value sensitive design, VSD, value tradeoffs, governance, policy, policy innovation, design psychology, applied ethics

**INTRODUCTION**

 The varied influences that artificial intelligence systems and robotics have on society have moved out of the realm of speculation and into reality. The impact that algorithmic trading agents, medical diagnostic systems, driverless cars and smart home assistants – to name a few – already have substantial and unignorable effects on the lives of both direct stakeholders (users, designers, companies, etc.) as well as indirect stakeholders (environments, bystanders, etc.). Their socialtechnicity – i.e., their inextricable link to the social environment in which they are designed and used – makes their study critical if their design and deployment are to be responsible. For this reason there has been a considerable amount of attention directed towards the ethical understanding of these systems, and a search towards actionable guidelines and best practices (Dignum, 2019). As a result, numerous principles, guidelines, recommendations, and values have been proposed to govern such systems, with a resulting risk of confusion as to which set to choose, thus delaying much-needed progress into making such principles actionable (Floridi et al., 2018). The next turn in AI ethics is how to translate abstract philosophical and legal principles/values into design requirements that engineers can understand and plan in design.

 Multiple approaches have emerged that consider the social embeddedness of technologies and their impacts. The core of many of these methodologies is the engagement and elicitation of stakeholders, whether they are directly or indirectly implicated by technology design. Approaches such as universal design (Ruzic & Sanfod, 2017; van den Hoven, 2017), inclusive design (Gregor, Sloan, & Newell, 2005; Hyppönen, Kemppainen, Gill, Slater, & Poulson, 2000), sustainable design (Fallan, 2015; Lockton, Harrison, & Stanton, 2016; Winkler & Spiekermann, 2019), participatory design (Bødker, Kensing, & Simonsen, 2009; Ehn, 2016), and value sensitive design (Friedman & Hendry, 2019; Umbrello, 2020a; van den Hoven & Manders-Huits, 2009) among others, have been constructed and proposed. Although these methodologies are disparate in many respects, they all aims towards the goal of responsible research and innovation (RRI).

 Originally developed within the field of human-computer interaction, value sensitive design (VSD) begins from the premise that technology is not value-neutral; rather, it is sensitive to stakeholder values, whether they are direct stakeholders such as users and designers, or indirect, such as the environment, and that social contexts and technologies co-vary (Friedman, Hendry, & Borning, 2017; van den Hoven & Manders-Huits, 2009). As a starting point then, the VSD approach aims to explicitly design technologies for stakeholder values – with emphasis on moral values - in a manner as to successfully map the values deemed critical and to ensure the robustness of sociotechnical systems (Friedman & Hendry, 2019; Umbrello, 2019b). What differs VSD from other design approaches then is its explicit emphasis on moral values and their inherit embeddedness in technologies (see Friedman & Hendry, 2019).

 VSD has traditionally prioritized the values that emphasize human well-being, human dignity, justice, welfare, and human rights as its central concern (Friedman, Kahn, Borning, & Huldtgren, 2013). The approach is considered ‘principled’ because it assumes an objective moral grounds on which these values spring, one that is independent of whether any particular individual or group subscribe to such values (e.g., the belief in and practice of racial eugenics by a group does not a priori mean that racial eugenics is a morally acceptable practice). Still, VSD maintains that expression of such values in any particular culture, or by any particular individual, can vary greatly (Friedman et al., 2017; Umbrello, 2020a). This ethical objectivism that VSD affirms easily permits it to be integrated into existent design practices across sociocultural dimensions, although it is not without objections (Davis & Nathan, 2014; Umbrello, 2018a, 2020).

 The ability for VSD to be adopted and integrated across sociocultural boundaries becomes invaluable, given the current calls for international collaboration and coordination for artificial intelligence regulations and policies. VSD has traditionally viewed policy as it would a technology. Accordingly, the approach would then need to identify policy requirements. Yet, the role of policy and how policy comes to play within VSD has not been seriously considered. This is exacerbated when we consider machine ethics policy that can have global consequences outside their development spheres. What constructs and models will position AI designers to engage in policy concerns? How can the design of AI policy be integrated with technical design? How might VSD be used to develop AI policy? How might law, regulations, social norms, and other kinds of policy regarding AI systems be engaged within value sensitive design? This chapter aims to touch on these fundamental questions and provide preliminary ways forward that can be useful to policy experts, AI researchers and designers, as well as academics.

 Previous studies have explored the philosophical basis of VSD (Santoni de Sio & van den Hoven, 2018; Umbrello, 2018b, 2020a), ad hoc applications of the approach to existent technologies (Correljé, Cuppen, Dignum, Pesch, & Taebi, 2015; van den Hoven, 2013; Woelfer, Iverson, Hendry, Friedman, & Gill, 2011), speculative applications of the methodology towards future technologies (Umbrello, 2019a; van Wynsberghe, 2013), as well as preliminary ways of incorporating VSD into AI design (Umbrello, 2019b; Umbrello & De Bellis, 2018). This chapter is comparatively unique, as it takes policy as the central object within VSD practice with specific emphasis on how machine ethics policies can emerge from VSD.

 To do this, this chapter is organized in five sections. The first section will look at how policy is typically engaged at the design level, looking specifically at AI research and development, as well as international development goals for which AI&R are situated. The second section briefly accounts the VSD approach in more detail, specifically looking at the tripartite structure of the approach and highlights where existent policy tools can come into play. Section three provides some preliminary suggestions on how the integrative nature of the approach allows for new policy recommendations to emerge. The fourth section discusses some of this chapter’s limitations, as well as provides suggestions for potentially fruitful future research streams.The final section concludes the chapter.

**POLICY, DESIGN AND COMPLIANCE**

` Policy formulation, development, and creation, like technologies, are motivated by problems in which successful policies are capable of solving those problems. Whether such policy measures are reactive, responding to problems that have arisen, or proactive, where policies are designed in anticipation of a problem that may arise, policy, regardless, is motivated by making things better than they are currently viewed (Simon, 1988).

 The traditional haphazard way of retroactively making policy is still widespread, and is not well-equipped to handle the exponential growth of technological innovations that consistently bring with them new social and ethical issues (Ben-Haim, Osteen, & Moffitt, 2013). Policy is typically understood as being mapped onto technological development during design from a top-down approach, ensuring client needs while also meeting the minimum standards of legal compliance to ensure safe deployment and minimize system recalcitrance (Bos, Walhout, Peine, & van Lente, 2014). Difficulties arise, however, when considering novel technologies such as artificial intelligence (AI) and robotics (AI&R), which exist across many domains, and in many cases overlap.

 Designing policies for AI&R that accurately and effectively permit interventions at a collective level becomes important given the potential consequences of recalcitrant systems. Given the number of abject policy failures (McConnell, 2010), unintended consequences, or ones that often produce the opposite results (Sieber, 2013), the goal towards collective policy to govern the design and use of AI&R poses a particularly exuberant challenge. Natural questions arise regarding the strategy towards policy creation, such as whether or not specific policy interventions should be developed for discrete technologies that risk being too narrow, and whether templates from narrow policies have any applicability to other domains. Similarly, can general policy measures that focus on principles or practices provide effective intervention strategies on the specific level?

 The two need not be read as a strong dichotomy however, mutually excluding each other. Design thinking shifts the view of policy as a static object, and it injects systems thinking as a fundamental principle of how policy co-varies and co-constitutes that which it seeks to police (Yoo et al., 2016). How existing policies can affect technological design, as well as how design can inform policy, becomes an interesting area to explore when we consider policy to be a similarly iterative and explorative enterprise, rather than as purely static process to which artifacts are forcefully subject.

**AI&R Policy**

 Currently, there exists no international policy or governance regulating artificial intelligence and robotics (Müller & Bostrom, 2016). Although broad and open discussions are currently being undertaken by various nation states, formalized guidelines remain varied, abstract, and mostly built upon existing policy and governance structures at the national or state levels. Aside from the always-present reasons why sovereign states fail to come to agreement as to the governance of any particular thing, one of the primary causes of this difficulty in global AI&R governance is the already-mentioned inability to agree upon the values upon which policy is built (Floridi et al., 2018; Umbrello & De Bellis, 2018). There is no explicitly adopted framework for engaging with such value tensions and moral overload (van den Hoven, Lokhorst, & van de Poel, 2012).

 Given that policy design is fundamentally a human endeavour – i.e., politically an endeavour of citizens – it necessarily implicates human values (Mintrom & Luetjens, 2017). Policy creation has typically taken a utilitarian/consequentialist approach to determining what would be most beneficial to stakeholders (Quah & Mishan, 2007). Cost-benefit analysis (CBA) is the primary framing of the utilitarian approach to policy analysis, and has been employed even in future technology speculation and ethics (Barrett & Baum, 2017; Umbrello & Baum, 2018). Although widely practiced, the CBA approach is fraught with inherent weaknesses, some that could prove particularly detrimental given the potential risks of recalcitrant AI&R systems (Baum, 2014; Muehlhauser & Bostrom, 2014; Soares & Fallenstein, 2014). CBA is predicated on the use of utility functions in which values are converted into monetary values, even those things which are traditionally considered irreducible to any numerical values such as art, happiness, and calmness among many others (Sunstein, 2013).

 When considering AI&R ethics, and AI&R policy more specifically, and considering the global effects that AI&R does and will continue to exhibit, it is important to account for a wider notion of stakeholder values that extends beyond the narrow confines of moral law theory approaches such as utilitarianism (Umbrello, 2020a). There is an increasing trend by citizens to resist policies that constrain fundamental values such as justice (Rawls, 2001), fairness/equality (Corak, 2016), and autonomy (Peters, 2018). Each of these values, among others, is critical to adoptable and effective policymaking, and because they are values commonly held as central to democratic systems, they thus bolster those systems. Yet they are each difficult, if not impossible, to quantify with monetary values as undertaken in the CBA approach.

**Democratic Commitments to AI&R Policy**

 We should resist being pessimistic as to the potential for global AI policy in which nation states unilaterally agree, though we should also confront the reality that such policy, albeit possible, may take many years, coming at the opportunity cost of addressing present and near-term concerns of AI&R that may otherwise be overlooked. A potentially fruitful initial first step, that is technically expounded in the proceeding sections, is beginning at national levels to design and deploy AI systems towards international goals (i.e., EU High Level Expert Group on Artificial Intelligence [HLEG AI], AI for Social Good factors [AI4SG]), most saliently the UN 2030 Sustainable Development Goals (SDGs).

 The United Nations proposed 2030 agenda for sustainable development was a set of goals directed at the design and implementation of a safe and sustainable future founded on the agreed desire for global peace (United Nations, 2019b). In 2015, all of the UN member states adopted the proposal, which aimed at the realization of 17 distinct, yet interconnected goals (See Figure 1). Part of the commitment of the SDG, which should be co-developed (i.e., not exclusive of one another), is a commitment to discussions on building strong and robust democratic institutions and governance structures [SDG #16] (United Nations, 2019a). A common element amongst different types of democratic regimes is a commitment to certain values, such as peaceful and inclusive societies for sustainable development, access to justice for all, as well as accountable and inclusive institutions at all levels (United Nations, 2019b).



*Figure 1. United Nations Sustainable Development Goals. Source:* (United Nations, 2019b)

The issues necessary for tackling these SDGs are varied, nuanced, and complex. To better address these issues holistically, the UN has a Technology Facilitation Mechanism (TFM) to encourage innovative solutions towards achieving the SDGs by adopting multi-stakeholder participation (United Nations, 2017, 2019). Before every UN session on SDGs, the TFM council assembles to discuss new solutions regarding the SDGs. The UN’s *modus operandi* with respect to the SDGs is that technology can be understood as both the problem and potential solution, as well as its adoption of the interactional stance towards technology. Instead of approaching technology as exclusively instrumental or as purely deterministic, it insists on the interactional performance between technology and social factors at an institutional level. By doing so, the UN aims to address these problems holistically, rather than in an ad hoc fashion.

 At some level then, SDGs can be understood as being the consequence of technologies, and that the more transformative a technology is, the greater its potential impact. High-speed trading algorithms, for example, provide asymmetric advantages to users, however they’re inaccessible to all but those who have the most capital already to employ such systems, which has the potential to exacerbate economic inequalities and unfair marketplaces (Busch, 2016). To this end, understanding the interactional nature of technology and social factors allows for more efficacious design and policy if tackled holistically. This means that AI systems, which are part of the larger sociotechnical *mise en scène* of information and communication technologies such as big data, deep learning, and cognitive computing, allow for a similar holistic approach towards SDG attainment. Umbrello and van de Poel (2020) already attempt to do this by using SDGs as a source of higher-level values that can then be translated through norms such as the AI4SG principles of Floridi, Cowls, King, and Taddeo (2020) to attain design requirements. This is done through the VSD method that is shown below. This approach, however, can be extended by framing it not in terms of specific design requirements for the technologies themselves, but for the policies that co-vary with those technologies. Similarly, because the values underpinning many of the SDGs are likewise fundamental to democratic regimes (a goal of the SDGs in itself), then it can be asserted that good AI&R policy is likewise aligned with such values; else it risks undermining itself.

 AI&R policy, for it to be efficacious, while similarly preserving the fundamental values of sustainable democracy, must then explicitly put democracy at its core (Ingram, deLeon, & Schneider, 2016). Value-sensitive policy design (VSPD), policy that is designed to be sensitive to these stated values, among others, can be viewed as co-constructive with democracy (i.e., it constructs democracy while also constructed by it). Given the various forms of democracy and the gradations of citizen participation and direct involvement, policy design will, like technology design, remain a dynamic practice, different between states (United Nations, 2019a). It can then be justified to state that any democratic regime engaging in AI&R policy design that is open to external information and engages with a wide stakeholder pool will consequently be more likely to produce AI&R policy that manifests a more expansive consideration of stakeholder concerns (both citizens and noncitizens).

 Because effective policy design is co-constructive of democracy, either supporting or constraining the values deemed important to the polity, it influences the level of democracy permitted within a nation. Given the stakes at play with technologies such as AI&R, which have global impacts, how to construct effective AI&R policy that manages the critical balance between individual autonomy (i.e., SDG #5) and security (i.e., SDG #16) – which is essential for either of the two to exist in any functional way (Bay, 1961) – becomes the central question of the design practice (Scheve & Stasavage, 2017). The following section introduces the VSD approach in greater detail, highlighting strengths of the approach that map onto the issues discussed, showing both how the VSD methodology is aptly suited to dissolving value conflicts and how it affirms the values central to democracy.

**VALUE SENSITIVE DESIGN – A DEMOCRATIC APPROACH TO AI&R DESIGN**

 As with all technologies that implicate some human values, the above analysis of policy similarly implicates human values. Meaning that, there will be inevitable value tensions that arise throughout the design phase as well as after its implementation. As is true traditionally with policy design, as well as with the currently unregulated AI R&D sphere, decisions about how an object is to be designed has typically rested in the hands of the designer (either the engineer, CEO, company, etc.). This top-down approach to design, regardless of the object of the design program, is bound to increase the probability of value tensions that can arise, given the technocratic way that designers go about evaluating the seemingly objective evidence on which the object is built.

 Policy, like technology, almost never arises *de nuovo*, but instead is built upon, or at least within, similar and familiar policy perspectives (Peters, 2018). This constrains designers to paradigms that might be inefficient in dealing with technologies like AI that cross national boundaries in which any single policy paradigm may dominate. This does not necessarily entail that good policy is deterministic, i.e., that because policy is constrained by previous policies and policy paradigms, the outcomes of newly introduced policies cannot be truly free from those constraints. Similarly, the technological *deterministic* approach rings a similar bell, in which the ‘inevitable’ march of progress cannot be halted, and that because each precursor technology determines the types of technologies that come after, humans are impotent to interfere with this progress (Woolgar, 1991).

 This deterministic perspective on either policies or technologies must be resisted if responsible innovation paradigms are to be actualized. Value sensitive design takes up an *interactional* stance on technological innovation design, arguing that technologies and societies are co-constructive of one another (Friedman & Hendry, 2019). Humans can guide the design of any given technology towards certain ends, and the technological impact and scope on society thus influence how we interact with that technology and each other (Friedman & Grudin, 1998). This perspective assigns technologies and policy a sufficient level of sociocultural influence while preserving the human potential to guide technologies, embedding them and designing them for the human values of stakeholders that are most beneficial. To that end, the design of socioculturally situated technologies can be achieved while simultaneously being geared towards more international ends by accounting for their impact of achieving the UN’s SDGs, which, of course, are intended to be globally beneficial.

**The Tripartite Methodology**

 The VSD approach has traditionally been described as a tripartite methodology consisting of three distinct, yet iterative *investigations*: 1) conceptual investigations, 2) empirical investigations, and 3) technical investigations (see Figure 2). Conceptual investigations involve designers consulting the philosophical literature that may be relevant to the technology under consideration in order to determine some prima facie values that may be implicated in the design program. It is during this point that designers can engage in preliminary stakeholder analyses to determine the relevant direct and indirect stakeholders that may be affected by the deployment of the technology (Borning & Muller, 2012). Empirical investigations take the conceptual work and elicit various stakeholder groups, enrolling them into the design program by employing social scientific means such as surveys, semi-structured interviews, envisioning cards, value sketches and scenarios (Friedman et al., 2017). The goal here is to better understand the values of stakeholders that may support or constrain the development of technology and vice versa. Finally, technical investigations look at how the architecture of the design object itself can support or constrain values.

*Figure 2. The recursive VSD tripartite framework employed in this study. Source:* (Umbrello, 2020b)

The approach can be broken down into at least eight steps that designers can follow. They are not to be taken as being in sequential order; instead, designers can begin with whichever step the design program calls for, increasing both the adoptability of the VSD approach into existent design domains, as well as increasing the overall design flow of a program.

1. **Begin by considering (1) a value, (2) a technology, or (3) the context of use**. VSD programs can begin with any one of the three. The ideal way to choose is whichever aligns most with the designer’s goals or interests (Figure 3).

*Figure 3. Starting considerations for VSD. Typically, one of the three is most pertinent to any given design. Source:* (Gazzaneo, Padovano, Umbrello, 2020)

1. **Direct and Indirect Stakeholders.** Methodically determine who the direct and indirect stakeholders are. Direct stakeholders are those individuals who interact directly with the technology itself or its output; indirect stakeholders instead are those individuals/groups that are also impacted by the system, though they never interact directly with it directly.
2. **Identify Harms and Benefits for Each Stakeholder Group.** Identify how the technology in questions will both positively and negatively affect each of the stakeholder groups.
3. **Map Harms and Benefits onto Corresponding Values.** At times the mapping between harms and benefits and corresponding values will be one of identity; at other times, the mapping will be multi-faceted (that is, a single harm might implicate multiple values, such as both security and autonomy).
4. **Conduct a Conceptual Investigation of Key Values.** Establish precise working definitions of each of the values elicited. Designers employ the philosophical literature in lieu of conceptual investigations to more accurately define these values and the potential issues that already exist with certain understandings of these values. How these values can be translated into norms and how those can then be translated into design requirements (and vice versa) is also investigated (See Figure 4).
5. **Identify Potential Value Conflicts.** In design spaces, value conflicts are typically not be read as ‘either/or’ dichotomies, but as tensions and constraints on the design space (van de Poel, 2014). Typical value conflicts include accountability vs. privacy, trust vs. security, environmental sustainability vs. economic development, privacy vs. security, and hierarchical control vs. democratization, among others (van den Hoven et al., 2012).
6. **Technical Investigation Heuristic and Value Conflicts.** Technical structures and tools will often dissolve, if not call, multiple conflicting values, often in the form of design trade-offs. Hence, designers here should aim to make explicit how a design trade-off maps onto a value conflict and differentially affects different groups of stakeholders (Umbrello, 2018b).
7. **Technical Investigation Heuristic and Unanticipated Consequences and Value Conflicts.** In order to be positioned to respond agilely to unanticipated consequences and value conflicts, when possible, design flexibility into the underlying technical architecture to support post-deployment modifications.

*Figure 4. Bi-directional values hierarchy (Source: Umbrello, 2019c)*

 The approach of VSD is then fundamentally participatory, democratically relying on the polity of stakeholders as a fundamental constituent of the design program that will ultimately impact them. Thus stakeholder elicitations, values, and analyses are what technologies, and perhaps policies, can be designed *for*; rather than remaining an afterthought, integrated *ex post facto* deployment, or overlooked entirely. There is, however, a certain level of technocracy within VSD. Although stakeholder values are ultimately the center of the design paradigm, the designers and design teams have final control of the design itself and the final product. Similar to representative democratic systems in which citizens participate but do not ultimately decide, VSD takes an approach of designing *for* citizen values, but designed by designers.

` We can begin to see how good policy – that which supports and in turn is supported by democratic regimes – is inextricably linked with VSD. Not only can the VSD approach be used as a democratic and participatory methodology by policymakers in drafting new policies for AI at a global level, but VSD is a form of policy in itself. Given the philosophical foundation of VSD as an interactional approach to considering technologies and their impacts on the implicated polity – and one that is fundamentally committed to certain universal values such as *human well-being, justice,* and *dignity* (Friedman & Hendry, 2019, p. 173) – VSD is political in that it affirms as its central values of considerations, those foundational to good policy and democracy: human well-being, human dignity, justice, welfare, autonomy, and human rights, among others (Friedman, Kahn Jr., & Borning, 2008). These are directly in line with the SDG’s outlines in the preceding sections, more saliently well-being as SDG #3, justice as SDG #16, and dignity as the culmination of all 17 SDGs.

**ILLUSTRATING VALUE SENSITIVE POLICY DESIGN AND VSD AS POLICY**

 As mentioned, VSD appears to be in possession of a double potency. Firstly, policy design can be accomplished through a VSD approach. Given that the approach affirms as its central values those that are necessary for the design of good policy; i.e., the values central to democracy and that are aligned with international goals towards those end since VSD is aptly capable as a design approach to achieving these ends (see Friedman & Kahn Jr., 2002). Secondly, because VSD is founded on the principles underlying those values, other objects of design, such as AI&R similarly, are brought into the paradigm of designing for democratic values. In this way, VSD supports democracy via design by acting as a policy-as-practice framework (see van de Kaa, Rezaei, Taebi, van de Poel, & Kizhakenath, 2019).

 Yet, these values remain purely abstract, even when considering an applied design approach such as VSD. To move from abstract to concrete, some examples for conceptualizing both of these potentialities that VSD possesses are presented below. Here some preliminary and cursory examples are given in the form presented in Figure 4 in which the bi-directional hierarchy of values can be conceptualized, either beginning with a value or the other way, beginning with a design requirement.

*Figure 5. Top-down approach beginning with the values of data privacy and consent that can then be translated into values through norms.*

Figure 5 is one of the ways in which two core values that have been heavily discussed in the field of informatics, particularly as it relates to machine learning and data set processing, are data privacy and consent (Floridi et al., 2020). VSD practitioners can translate the values through various sociocultural norms in which the technology program is situated in order to more accurately determine the technical design requirements that can map those values. The norms, and even specific design requirements, can be policy-driven if existent. For example, the EU’s General Data Protection Regulation 2016/679 (GDPR) provides norms that various actors must comply to, with regards to the processing and transfer for personal data. The norm ‘maximize privacy,’ for example, can be satisfied under the GDPR if the systems employed in design are intelligible to data protection officers to ensure compliance of the processing and use of data in accordance with the other statues laid out in the GDPR. This is one of a myriad of ways of conceptualizing such values as design requirements; the visualization is employed by designers to better conceptualize how to plan design programs to be compliant to stakeholder values, the GDPR being one of a multitude of considerations.

 Data privacy and consent are taken as examples of the many values that are considered when designing AI&R, however, it is particularly salient for this discussion given their importance to other democratic values such as safety and human dignity, which they could not exist without (Nissenbaum, 2009; Solove, 2008). Because personal information is often considered as foundational to the concept of the individual, the core of democracy, respect for data privacy, and consent of its use, are similarly the core of affirming individual safety and dignity (Floridi, 2016). Similarly, given that data privacy forms one of the foundational values of the VSD approach (Umbrello, 2019b; Umbrello & De Bellis, 2018), its adoption for the development of AI&R systems proves particularly relevant and salient.

 As mentioned, however, the VSD approach need not only be seen as a tool for designing technologies for existent policies, but it is also a form of policy itself, given the fundamental values on which the approach has been built are those of good policy; e.g., democratic values. If we take, for example, the same hierarchy in Figure 4, but reverse it – solely for illustration’s sake of working from the bottom up – we can affirm certain values *ex post facto*, beginning with design requirements (see Figure 6).

*Figure 6. Beginning with design requirements that can uncover and inform policy requirements*

 Certain design requirements for technologies under question may be presented *prima facia*. The VSD approach is aptly equipped to begin from this stage, translating these design requirements into more general sociocultural norms (such as the AI4SG principles outlined by Floridi et al. 2020) and finally into more universal moral values (i.e., UN SDGs). Figure 6 illustrates how certain design requirements can be translated into norms that can satisfy a functional definition of transparency or explicability within the context of algorithmic decision-makers. They are in no way exclusive nor exhaustive, but demonstrate home some values that may be important, if not fundamental to certain technologies, in this case AI, are crucial if they are to be development for the social good in mind (Floridi et al., 2020). As such, the VSD approach, adopted within the general and distributed context of designing AI&R for social good, can be used to help inform policymakers in developing for certain stakeholders the values deemed necessary for those ends.

 This becomes particularly prescient when considering the emerging sociocultural and ethical impacts of transformative technologies such as AI&R will have as they become more embedded in different sociocultural contexts. VSD’s fundamental practice of being iterative and self-improving situates it in a unique position to be adopted by engineers and designers to be compliant to stakeholder values while simultaneously helping policymakers to make more technologically compatible policy decisions that support responsible innovation, both at national and international levels.

**LIMITATIONS AND FUTURE RESEARCH PROJECTS**

 The aim of this chapter has been to consider the design of good policy, particularly that of policy regarding artificial intelligence and robotics within the value sensitive design approach. How VSD can be used to design AI&R to be compliant to stakeholder values and current policy has been shown, and historically has marked VSD as being an aptly suited approach in doing so (van Wynsberghe & Robbins, 2014; Warnier, Dechesne, & Brazier, 2014; Woelfer et al., 2011). Similarly, how values can be derived during the design process when beginning with the technology has been discussed, and how such can be used to help policymakers make more relevant and consistent decisions when considering AI&R.

 Still, there are areas of limitations that this chapter does not explore, but warrant future research. Firstly, researchers should look at making the VSD explicitly functional as being a policymaking tool, one that policymakers themselves look to, for making intelligible technology policy. Similarly, and perhaps more pressing, is how policymakers can make VSD approaches more adoptable by current AI&R groups to create a more homogenous landscape in which compliance can be adjudicated, and how emerging and new values and policy recommendations can be explicated.

 Ultimately, what is needed are real-world examples of VSD being employed as an explicit policy tool to better determine its applicability as such. How VSD can handle different policies for different sociocultural contextualized versions of AI&R is yet to be seen, but critical in determining how VSD can be used to create good policy that supports and balances the stakeholder values that are central to democratic regimes. Nonetheless, the need for open and collaborative policymaking is required if AI&R is to be designed in a responsible way that supports the values that stakeholders affirm. This chapter seeks to spark the conversation on how to put these values into practice, given the urgency of developing a transformative technology that is already underway.

**CONCLUSIONS**

 This chapter looked at how the VSD methodology is aptly suited as a design approach to designing good policy. VSD’s core foundation is predicated on the values that support those of democratic regimes, which places it in a unique position to design policy as it would any other technology, by enrolling stakeholders, eliciting their values, mapping harms and benefits, and designing policy that is interactive, modifiable, and reflective of those values. The specific example of AI&R is used to illustrate the salience and necessity of adopting such an approach if collective policy and governance of those systems is to be achieved. Similarly, an argument was also put forward that VSD, given its philosophical foundations ahd starting points, is itself a form of policy as it guides design toward certain political ends, those that are affirmed by democratic regimes.

**AKNOWLEDGMENTS**

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. All remaining errors are the author’s alone. The views expressed are not necessarily those of the Institute for Ethics and Emerging Technologies. The content of this publication has not been approved by the United Nations and does not reflect the views of the United Nations or its officials or Member States.

**REFERENCES**

Barrett, A. M., & Baum, S. D. (2017). Risk Analysis and Risk Management for the Artificial Superintelligence Research and Development Process. In V. Callaghan, J. Miller, R. Yampolskiy, & S. Armstrong (Eds.), *The Technological Singularity: Managing the Journey* (pp. 127–140). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-54033-6\_6

Baum, S. D. (2014). The great downside dilemma for risky emerging technologies. *Physica Scripta*, *89*(12), 10. https://doi.org/10.1088/0031-8949/89/12/128004

Bay, C. (1961). The Structure of Freedom. *Science and Society*, *25*(1).

Ben-Haim, Y., Osteen, C. D., & Moffitt, L. J. (2013). Policy dilemma of innovation: An info-gap approach. *Ecological Economics*, *85*, 130–138. https://doi.org/10.1016/j.ecolecon.2012.08.011

Bødker, K., Kensing, F., & Simonsen, J. (2009). *Participatory IT design: Designing for business and workplace realities*. MIT press.

Borning, A., & Muller, M. (2012). Next steps for value sensitive design. *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI ’12*, 1125. https://doi.org/10.1145/2207676.2208560

Bos, C., Walhout, B., Peine, A., & van Lente, H. (2014). Steering with big words: articulating ideographs in research programs. *Journal of Responsible Innovation*, *1*(2), 151–170. https://doi.org/10.1080/23299460.2014.922732

Busch, D. (2016). MiFID II: regulating high frequency trading, other forms of algorithmic trading and direct electronic market access. *Law and Financial Markets Review*, *10*(2), 72–82.

Corak, M. (2016). ‘Inequality is the root of social evil,’or Maybe Not? Two Stories about Inequality and Public Policy. *Canadian Public Policy*, *42*(4), 367–414.

Correljé, A., Cuppen, E., Dignum, M., Pesch, U., & Taebi, B. (2015). Responsible Innovation in Energy Projects: Values in the Design of Technologies, Institutions and Stakeholder Interactions. In B.-J. Koops, I. Oosterlaken, H. Romijn, T. Swierstra, & J. van den Hoven (Eds.), *Responsible Innovation 2* (pp. 183–200). Springer International Publishing. https://link.springer.com/chapter/10.1007%2F978-3-319-17308-5\_10

Davis, J., & Nathan, L. P. (2014). Value Sensitive Design: Applications, Adaptations, and Critiques. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* (pp. 1–26). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6994-6\_3-1

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer International Publishing. https://doi.org/10.1007/978-3-030-30371-6

Ehn, P. (2016). Design, Democracy and Work: Exploring the Scandinavian Participatory Design Tradition.

Fallan, K. (2015). Our Common Future. Joining Forces for Histories of Sustainable Design. *TECNOSCIENZA: Italian Journal of Science & Technology Studies*, *5*(2), 15–32. http://www.tecnoscienza.net/index.php/tsj/article/view/201%5Cnhttp://www.tecnoscienza.net/index.php/tsj/article/download/201/136%5Cnhttp://www.tecnoscienza.net/index.php/tsj/article/view/201/136?utm\_source=non+iscritti+a+TS&utm\_campaign=e868895d7d-newissu

Floridi, L. (2016). On human dignity as a foundation for the right to privacy. *Philosophy & Technology*, *29*(4), 307–312.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(December), 1–24. https://doi.org/10.12932/AP0443.32.4.2014

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). Designing AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*, 1–26. https://doi.org/10.1007/s11948-020-00213-5

Friedman, B., & Grudin, J. (1998). Trust and accountability: preserving human values in interactional experience. In *CHI 98 conference summary on Human factors in computing systems* (p. 213). ACM.

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. Cambridge, MA: MIT Press.

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

Friedman, B., & Kahn Jr., P. H. (2002). Value sensitive design: Theory and methods. *University of Washington Technical*, (December), 1–8. https://doi.org/10.1016/j.neuropharm.2007.08.009

Friedman, B., Kahn Jr., P. H., & Borning, A. (2008). Value Sensitive Design and Information Systems. *Human-Computer Interaction and Management Information Systems: Foundations*, 69–101. https://doi.org/10.1145/242485.242493

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value Sensitive Design and Information Systems. In N. Doorn, D. Schuurbiers, I. van de Poel, & M. E. Gorman (Eds.), *Early engagement and new technologies: Opening up the laboratory* (pp. 55–95). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-7844-3\_4

Gazzaneo, L., Padovano, A., & Umbrello, S. (2020). Designing Smart Operator 4.0 for Human Values: A Value Sensitive Design Approach. In *International Conference on Industry 4.0 and Smart Manufacturing (ISM 2019) in Procedia Manufacturing* (pp. 219–226). Rende, CS: Elsevier. https://doi.org/10.1016/j.promfg.2020.02.073

Gregor, P., Sloan, D., & Newell, A. F. (2005). Disability and Technology: Building Barriers or Creating Opportunities? *Advances in Computers*. https://doi.org/10.1016/S0065-2458(04)64007-1

Hyppönen, H., Kemppainen, E., Gill, J., Slater, J., & Poulson, D. (2000). *Handbook on Inclusive Design of Telematics Applications.* (H. Hyppönen, E. Kemppainen, J. Gill, J. Slater, & D. Poulson, Eds.). Themes/STAKES 2.

Ingram, H., deLeon, P., & Schneider, A. (2016). Conclusion: Public policy theory and democracy: The elephant in the corner. In *Contemporary Approaches to Public Policy* (pp. 175–200). Springer.

Lockton, D., Harrison, D., & Stanton, N. A. (2016). Design for Sustainable Behaviour: investigating design methods for influencing user behaviour. *Annual Review of Policy Design*, *4*(1), 1–10.

McConnell, A. (2010). *Understanding policy success: Rethinking public policy*. Macmillan International Higher Education.

Mintrom, M., & Luetjens, J. (2017). Creating public value: Tightening connections between policy design and public management. *Policy Studies Journal*, *45*(1), 170–190.

Muehlhauser, L., & Bostrom, N. (2014). Why We Need Friendly AI. *Think*, *13*(36), 41–47. https://doi.org/10.1017/S1477175613000316

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion BT - Fundamental Issues of Artificial Intelligence. In V. C. Müller (Ed.) (pp. 555–572). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1\_33

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Peters, B. G. (2018). *Policy problems and policy design*. Edward Elgar Publishing.

Quah, E., & Mishan, E. J. (2007). *Cost-benefit analysis*. Routledge.

Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press.

Ruzic, L., & Sanfod, J. A. (2017). Universal Design Mobile Interface Guidelines (UDMIG) for an Aging Population. In *Mobile e-Health* (pp. 17–37). Springer.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*. https://www.frontiersin.org/article/10.3389/frobt.2018.00015

Scheve, K., & Stasavage, D. (2017). Wealth inequality and democracy. *Annual Review of Political Science*, *20*, 451–468.

Sieber, S. (2013). *Fatal remedies: The ironies of social intervention*. Springer Science & Business Media.

Simon, H. A. (1988). The science of design: Creating the artificial. *Design Issues*, 67–82.

Soares, N., & Fallenstein, B. (2014). *Agent Foundations for Aligning Machine Intelligence with Human Interests : A Technical Research Agenda*. https://doi.org/10.1007/978-3-662-54033-6\_5

Solove, D. J. (2008). *Understanding privacy* (Vol. 173). Harvard university press Cambridge, MA.

Sunstein, C. R. (2013). The value of a statistical life: some clarifications and puzzles. *Journal of Benefit-Cost Analysis*, *4*(2), 237–261.

Umbrello, S. (2018a). *Safe-(for whom?)-by-Design: Adopting a Posthumanist Ethics for Technology Design*. York University. https://doi.org/10.13140/RG.2.2.29726.38720

Umbrello, S. (2018b). 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

Umbrello, S. (2019a). 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

Umbrello, S. (2019b). Beneficial Artificial Intelligence Coordination by Means of a Value Sensitive Design Approach. *Big Data and Cognitive Computing*, *3*(1), 5. https://doi.org/10.3390/bdcc3010005

Umbrello, S. (2020a). Imaginative Value Sensitive Design: Using Moral Imagination Theory to Inform Responsible Technology Design. *Science and Engineering Ethics*, *26*(2), 575–595. https://doi.org/10.1007/s11948-019-00104-4

Umbrello, S. (2020b). Meaningful Human Control Over Smart Home Systems: A Value Sensitive Design Approach. *HUMANA.MENTE Journal of Philosophical Studies*, *13*(37), 40-65. https://www.humanamente.eu/index.php/HM/article/view/315

Umbrello, S., & Baum, S. D. (2018). Evaluating future nanotechnology: The net societal impacts of atomically precise manufacturing. *Futures*, *100*(June), 63–73. https://doi.org/10.1016/j.futures.2018.04.007

Umbrello, S., & De Bellis, A. F. (2018). A Value-Sensitive Design Approach to Intelligent Agents. In R. V. Yampolskiy (Ed.), *Artificial Intelligence Safety and Security* (pp. 395–410). CRC Press. https://doi.org/10.13140/RG.2.2.17162.77762

Umbrello, S., & van de Poel, I. (2020). *Mapping Value Sensitive Design onto AI for Social Good Principles*. https://www.academia.edu/43347384/Mapping\_Value\_Sensitive\_Design\_onto\_AI\_for\_Social\_Good\_Principles

United Nations. (2017). *Data Privacy, Ethics and Protection Guidance Note on Big Data For Achievement of the 2030 Agenda.* https://unsdg.un.org/resources/data-privacy-ethics-and-protection-guidance-note-big-data-achievement-2030-agenda

United Nations. (2019a). Democracy & the SDGs. https://sustainabledevelopment.un.org/index.php?page=view&type=20000&nr=5780&menu=2993

United Nations. (2019b). Sustainable development goals. *GAIA*. https://doi.org/10.14512/gaia.28.2.1

van de Kaa, G., Rezaei, J., Taebi, B., van de Poel, I., & Kizhakenath, A. (2019). How to Weigh Values in Value Sensitive Design: A Best Worst Method Approach for the Case of Smart Metering. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-019-00105-3

van de Poel, I. (2014). Conflicting Values in Design. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* (pp. 1–23). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6994-6\_5-1

van den Hoven, J. (2013). Architecture and Value-Sensitive Design. In C. Basta & S. Moroni (Eds.), *Ethics, design and planning of the built environment* (p. 224). Springer Science & Business Media. https://books.google.ca/books?id=VVM\_AAAAQBAJ&dq=moral+value+such+as+freedom,+equality,+trust,+autonomy+or+privacy+justice+%5Bthat%5D+is+facilitated+or+constrained+by+technology&source=gbs\_navlinks\_s

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, UK: Cambridge University Press. https://doi.org/10.1017/9780511844317

van den Hoven, J., Lokhorst, G. J., & van de Poel, I. (2012). Engineering and the Problem of Moral Overload. *Science and Engineering Ethics*, *18*(1), 143–155. https://doi.org/10.1007/s11948-011-9277-z

van den Hoven, J., & Manders-Huits, N. (2009). Value-Sensitive Design. In *A Companion to the Philosophy of Technology* (pp. 477–480). Wiley-Blackwell. https://doi.org/10.1002/9781444310795.ch86

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

van Wynsberghe, A., & Robbins, S. (2014). Ethicist as Designer: A Pragmatic Approach to Ethics in the Lab. *Science and Engineering Ethics*, *20*(4), 947–961. https://doi.org/10.1007/s11948-013-9498-4

Warnier, M., Dechesne, F., & Brazier, F. (2014). Design for the Value of Privacy. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* (pp. 1–14). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6994-6\_17-1

Winkler, T., & Spiekermann, S. (2019). Human Values as the Basis for Sustainable Information System Design. *IEEE Technology and Society Magazine*, *38*(3), 34–43. https://doi.org/10.1109/MTS.2019.2930268

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. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1707–1716). New York, NY, USA: ACM. https://doi.org/10.1145/1978942.1979191

Woolgar, S. (1991). The Turn to Technology in Social Studies of Science. *Science, Technology, & Human Values*, *16*(1), 20–50. https://doi.org/10.1177/016224399101600102

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. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4423–4434). ACM.