The moral psychology of Value Sensitive Design: the methodological issues of moral intuitions for responsible innovation

Preprint · January 2018

DOI: 10.13140/RG.2.2.31213.69605

1 author:

Steven Umbrello
Institute of Ethics and Emerging Technologies
40 PUBLICATIONS **0** CITATIONS

The moral psychology of Value Sensitive Design: the methodological issues of moral intuitions for responsible innovation

Steven Umbrello[a]

[a] Institute for Ethics and Emerging Technologies

ARTICLE HISTORY
Compiled January 20, 2018

ABSTRACT
This paper argues that although moral intuitions are insufficient for making judgments on new technological innovations, they maintain great utility for informing responsible innovation. To do this, this paper employs the Value Sensitive Design (VSD) methodology as an illustrative example of how stakeholder values can be better distilled to inform responsible innovation. Further, it is argued that moral intuitions are necessary for determining stakeholder values required for the design of responsible technologies. This argument is supported by the claim that the moral intuitions of stakeholders allow designers to conceptualize stakeholder values and incorporate them into the early phases of design. It is concluded that design-for-values (DFV) frameworks like the VSD methodology can remain potent if developers adopt heuristic tools to diminish the influence of cognitive biases thus strengthening the reliability of moral intuitions.

KEYWORDS
design methodology; moral epistemology; design psychology; innovation; value sensitive design

1. Introduction

It has long been a contention in the sociology of science that technological artifacts are embedded with values, whether those values are explicitly designed into them or not (Magnani, 2013; Pinch & Bijker, 1987; van Wynsberghe & Robbins, 2014; Winner, 2003). As such, technological artifacts become the subject of ethical discourses as the values that are embedded become of political and social import. The issues associated with the values in design of technologies becomes exacerbated when we consider the transformative nature of emerging technologies (King, Whitaker, & Jones, 2011; Lucivero, Swierstra, & Boenink, 2011; Roache, 2008; Timmermans, Zhao, & van den Hoven, 2011). This paper shows how many of the design judgments made regarding emerging technologies are the result not of ethical deliberation but instead of moral intuition. This moral intuition is shown to be insufficient to responsibly inform research and innovation.

Since the 1970's, the study of moral psychology has illuminated not only the ways in which individuals think about and use their faculty of moral intuitions but also the extent to which those intuitions map onto the real world usage of technology. Addition-

E-mail: steve@ieet.org

ally, empirical work in psychology has shown that a number of cognitive biases may in fact influence individuals' reasoning processes (e.g., Caviola, Mannino, Savulescu, & Faulmuller, 2014; Tversky & Kahneman, 1974). This influence and the resultant biased reasoning in moral intuition is particularly evident when the subject of intuition is unique, convoluted or laden with ideology, such as technological innovations are (Caviola et al., 2014; Cosmides & Tooby, 1992; Kahan, Peters, Dawson, & Slovic, 2013).

It is the purpose of this paper to outline how technological innovations reveal the limits of moral intuitions in such a way as to make them apparently insufficient in informing responsible innovation (RI). In response to such impetus, I argue that it is *only* via the inclusion of moral values that RI, viz. value integration at the early design phases, is possible. To this end, I draw upon the Value Sensitive Design (VSD) approach amongst the existent design-for-values approaches (DFV), to illustrate this thesis. VSD is chosen in particular amongst other design methodologies such as participatory design, universal design and inclusive design particularly because it both mandates that designers account for the *explicit* values of stakeholders as well as how those values map onto the existent ethical literature. Thus, showing how this philosophically informed framework may be contested has foundational implications for all DFV approaches which hinge on the incorporation of stakeholder values in technological design. Given that the proposed VSD approach requires designers to investigate the needs and values of stakeholders, the moral intuition of those stakeholders becomes of particular importance. Hence, given arguments against the favourability of moral intuitions towards novel technologies, this paper concludes by claiming that the DFV frameworks can, in fact, remain successful methodologies if designers approach their investigations with a toolkit of simple heuristics that can reduce the influence of bias on moral intuition.

To the best of my knowledge, this paper is the first to: 1) evaluate the merits of a DFV approach from the perspective of moral, psychological theory, and 2) restructure such an approach in such a way as to make it more adaptable to empirical evidence stemming from moral psychology. Prior literature on VSD has focused on methodology (Cummings, 2006; Friedman & Kahn Jr., 2002; Friedman, Kahn Jr., Borning, & Huldtgren, 2013; Van den Hoven, Lokhorst, & Van de Poel, 2012; van den Hoven & Weckert, 2008), applications to current technological innovations (Aad Correlje, Eefje Cuppen, Marloes Dignum, 2015; Briggs & Thomas, 2015; van den Hoven, 2007) as well as to novel technologies (Dechesne, Warnier, & van den Hoven, 2013; Friedman, 1997; Friedman & Kahn Jr., 2000; Timmermans et al., 2011; van den Hoven, 2014; van Wynsberghe, 2013; Van Wynsberghe, 2016). Although these studies provide useful information, they do not take into account the reliability of all of the constituent parts of the VSD approach, in this case, the importance of moral intuition in conceptual investigations. This paper's application of moral intuition and cognitive biases towards converging technologies and VSD is comparatively unique. It is the intent of this paper to help spur continued discussions on some of the most pertinent ethical issues regarding the design of emerging and converging technologies.

To successfully tackle these arguments, this article is organized into the following sections: the first section will give a more in-depth account of 1) how moral intuitions are commonly understood in ethics and moral psychology, 2) the strength of moral intuitions when applied to novel technologies, in particular, nano-bio-info-cogno (NBIC) technologies. The second section will lay out the methodological framework of the VSD approach so that the reader can better understand the position that values, and consequentially, moral intuitions have in the theory. The third section will draw upon the

conclusions of the first section to argue that the VSD approach needs to adopt heuristic tools to strengthen its conceptual investigations in determining stakeholder values. The final section of this paper sketches the broader theoretical implications that these conclusions induce as well as potential fruitful research streams.

2. Moral intuitiveness and cognitive biases

The application of moral concepts such as deontology (i.e. duty ethics), utility ethics (e.g. utilitarianism, consequentialism) or virtue ethics definitely have positions in conversations about the way that we should act as it concerns emerging technologies. However, this paper's primary focus concerns the application and utility of moral intuitions, fraught as they are. In order to do this, I devote the rest of this section to discussing current conceptions of moral intuitiveness as well as some of the barriers that moral judgments inevitably encounter.

Although the early history of moral psychology was primarily dominated by rationalist theories of morality, recent decades have seen a shift, given new empirical findings and advances in psychology (Haidt, 2001). This change has led to the adoption and exploration of what is now known as Intuitionism. Moral intuition is thus far removed from the theoretical underpinnings of rationalist theory given that intuitionism is argued to be a process of cognition rather than a rationalization in search of moral truths that then inform moral judgments. Intuitionists argue that rationalist conception is instead a mischaracterization of actual occurrences in the brain; individuals first intuit moral judgments, then only on an ex post facto basis does the agent rationalize their decision (Haidt, 2001; Shweder & Haidt, 1993). Mostly, agents make automatic value judgments, typically unaware of the cognitive processes that produced such judgments. It is only after they are pressed for reasons that agents attempt to give an argumentative rational for why they arrived at such a judgment, sometimes unconvincingly (Haidt, 2001). Haidt (2001) constructs a narrative of safe, consensual incest. This narrative invokes a visceral sense of 'wrongness' that agents who are presented this story are hard pressed to give reasons for why they find it morally abhorrent. Likewise, Klein (2016) proposes a variation of the trolley problem that also elicits a 'wrongness' feeling. Both examples are cited to illustrate how moral intuition works. Individuals are presented cases; they are, in turn, asked to make a value judgment, then once pressed for reasons for such judgement, they find difficulty in rationalizing the inclination, thus emphasizing a lack of a priori reasoning (Klein, 2016).

Nonetheless, whether or not we take moral intuitionism as a real interpretation of how human moral judgments are formed, issues still arise that provide practical problematic concerns. These problems are typically manifested in how we apply our moral judgments as a product of intuition, primarily when intuition has been shown to be highly susceptible to a number of cognitive biases (Brink, 2014; Cushman, Young, & Hauser, 2006; Greene, 2014; Nichols & Knobe, 2007; Waldmann & Dieterich, 2007; Woodward & Allman, 2007). Such biases can affect the ways in which we value technologies, the way we approach technologies and the way in which we interact with them. Missteps can cause a slowdown of progress at best and increase the risks of catastrophic events at worst (Bostrom, 2014; Caviola et al., 2014).

In fact, Caviola et al. (2014) explain the potential effects of cognitive biases on the value-perception of cognitive enhancing technologies (CE). They argue that the polarized nature of the debates surrounding the ethical issues of CE can be explained by the influence of biases on human reasoning, thus disposing the moral intuition of

individuals to intuit in particular ways. They conclude that not only do cognitive biases have the potential to lead people to make irrational value-judgments about CE but that they are more likely to do so in a negative capacity. Likewise, these biases are not exclusive to debates regarding CE but may be just as pervasive in discussions of other transformative technologies.

Klein (2016) takes a different approach to our application of moral intuitions to novel technologies. He argues that given our evolutionary history, and the rate of technological innovation, humans have failed to acquire moral intuitions in a capacity that can sufficiently make value-judgments of technology. Because the increasing complexity of these technologies makes the causal chains ambiguous, he argues that we naturally tend to miss or ignore the ethical issues that emerge with these technologies. As a consequence, Klein proposes that moral intuitions, because of their innate failures, must be buttressed with "culture substitutes" that can help to make complex and novel technologies easier to intuit. Because he argues that we are good at making moral intuitions about our fellow human behaviour, one way to substitute intuition is by employing what he calls a "what if it was human" method. This process involves conceptualizing novel technology as if its uniqueness was embodied by a person. In speculating how that person would behave, both publically and privately, can assist in intuiting the moral value of the technology.

Regardless, the importance of moral intuition as an essential way of understanding human judgment has become a dominating area of discussion in moral psychology. The shift away from the traditional humanist view of humans as rational animals towards one that argues that most of our moral judgments are a product of the unconscious has significant implications for how we view emerging and converging technologies; particularly given the susceptibility of moral intuitions to biases (Klein 2016).

Taking this into account, it is this paper's contention is that the DFV approaches, particularly VSD, although claiming to be predicated on a foundation of universal moral values is in fact not, instead, upon closer inspection, most of the values incorporated are intuition based. Not only this, but I contend that objectivity in the determination of values is not necessary. A functionalist, pragmatic approach is thus adopted for the purposes of this paper. The intersubjectivity of moral intuitions, the acknowledgment of this in DFV approaches and the strengthening of moral intuitions through heuristic tools as mentioned in this paper provides at least one means by which we can reflexively engage in RRI. The

As a consequence, any design framework or methodology that works as a function of stakeholder values must take into account the inherent susceptibility of moral intuitions to cognitive biases. The following section introduces the design methodology VSD. As noted earlier, this paper aims to bolster the applicability of the VSD approach towards the responsible innovation of transformative technologies. Because the discussion of values in this paper is not restricted to VSD (meaning that the values and the related issues are not VSD-exclusive), the implications drawn from this paper have a scope that extends beyond VSD.

3. Value Sensitive Design (VSD)

Value Sensitive Design is an approach to the design of technology that is unique in its methodology of accounting for human values during the design phases (Friedman & Kahn Jr., 2002). Originating in the domain of Human-Computer Interaction (HCI), VSD has since been developed as a proposed approach to the responsible innovation

of many different technologies such as identity technologies (Briggs & Thomas, 2015), energy technologies (Aad Correlje, Eefje Cuppen, Marloes Dignum, 2015), information and communication technology (ICT) (Dechesne et al., 2013; Friedman, 1997; Friedman et al., 2013; Huldtgren, 2014; van den Hoven, 2007) and nanotechnology (Timmermans et al., 2011; van den Hoven, 2014).

VSD is one of a host of 'design-for-values' or 'safe-by-design' approaches to RI (see Micheletti & Benetti, 2016). The inception of these methodologies are in response to some of the foundational issues debated within RI discourses, primarily those that result from the the social construction and co-production of technological artifacts (Foley, Bernstein, & Wiek, 2016; see also Pinch & Bijker, 1987). Firstly, the issues associated with the infrastructural embeddedness of a technology over a period of time make modification difficult. Hence, once the negative impacts of a technology emerge, they may have already enrolled economic and political capital that make augmentations challenging (Collingridge, 1980; Star, 1999). As such, we see how the governance of technological artifacts are ex post facto, meaning they are retrospective in nature and usually follow ubiquitous production of technologies (Kaiser, Kurath, Maasen, & Rehmann-Sutter, 2009; Rip, 2009; Rip & Van Amerom, 2009). Likewise, there is a seeming gap between the the enterprises of technological development and societal requirements (Daniel, 2002; Sarewitz & Pielke, 2007).

DFV frameworks such as VSD were developed in the attempt to address these challenges. They are are attempts at balancing often divergent approaches to the amelioration of these challenges, primary between risk-based and precautionary approaches to design (Alvial-Palavicino, 2016; Brey, 2012; Brown, 2009; Guston, 2014; Nordmann, 2014; te Kulve & Rip, 2011). In doing so they generally aim to address these issues via design by making foundational the dimensions of anticipation (future values, issues), reflexivity (biases, assumptions, intentions), inclusion (of various stakeholders) and responsiveness (recursively operationalizing the former three practice) (Owen et al., 2013; Owen, Macnaghten, & Stilgoe, 2012). Each, however, operationalize these dimensions in their own way. This paper employs the VSD methodology because its formalization of these dimensions is overtly explicit, making its evaluation of value-inclusions simpler to demonstrate and its broader implications less convoluted.

VSD is defined as "a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" (Friedman & Kahn Jr., 2002, 1). VSD is grounded on the foundational premise that technologies embody values (they are value-laden) and provides a framework and methodology for assessing the current design of technologies while simultaneously integrating a proactive approach to guide the development of technologies both at the early stages of design and throughout the design process. What differentiates VSD from other DFV approaches are seven distinct characteristics in conjunction with one another:

(1) VSD aims to direct the development of technology not only through manipulations in the process of design and development but in the early stages of design.
(2) VSD does not only incorporate the values of designers or those directly involved as stakeholders, but also the public at large, industry and other sectors.
(3) VSD does not focus solely on the values gained through participatory or democratized means but rather seeks to account for all relevant values with particular weight given to values with moral weight.
(4) VSD is a harmonized methodology that consists of three separate, but integrated types of analyses: conceptual, empirical, and technical.

(5) VSD does not view values as something that arise out of necessity from the technology nor do values passively come from societal forces. Instead, VSD is 'interactional,' meaning that values are dynamic as technology affects individual behaviour and society affects technological progress.
(6) VSD gives particular weight to the values of justice, human rights, and human welfare. These values, taken from moral epistemology, are understood as being evaluated independently of subjective or culture belief in them.
(7) Beyond moral epistemology, VSD methodology assumes that upon close analysis of values, some values can be determined to be universal between differing cultures and societies. Although these values may manifest themselves in varying ways, upon close analysis one can decide that in fact those manifestations are simply a variation of a universal value. Examples include freedom, trust, equality, and privacy

All in all, VSD is a methodology that has been designed with the intention of being capable of integrating the values of stakeholders during the early design stages to guide the technology in a more predictable way, while still allowing for the flexibility to account for emerging changes in values and impacts (Fig. 1). Those individuals and groups that interact directly with the technology in question are considered direct stakeholders while those in the peripheries are considered indirect. The VSD approach requires that designers account for both direct and indirect stakeholder values in the design phases, the latter of which is typically side-lined during conventional design processes (Friedman & Kahn Jr., 2002; Taebi, Correljé, Cuppen, Dignum, & Pesch, 2014). This is not only done via consultation with existing literature on what stakeholders value, bit through their direct enrollment. This means that diverse publics and differing epistemic patches are particularly levied in order to more richly enhance the legitimacy and salience of design (Cash et al., 2003; Chilvers, 2007; Delgado, Kjølberg, & Wickson, 2010).

Additionally, one of the primary concerns of VSD is that issues arise with the application of technology as a result of the ethical values that society holds related to that application (Timmermans et al., 2011). In acknowledging this conflict, VSD aims to account for the relevant societal values and address the potential conflict during design stages (Capurro, Longstaff, Hanney, & Secko, 2015; Friedman & Kahn Jr., 2002; Taebi et al., 2014). This means not only designing *in* what values are determined most pertinent, but also designing *out* any unwanted values. This involves an awareness that designers can implicitly embed values given their relatively centralized positions in the design process.

A step-by-step methodology for implementing the VSD framework has already explicated by Friedman, Kahn and Borinng (2008). As such I have chosen to forgo a rephrasing of it given that the intention of this paper is not to advocate for the VSD framework over other DFV methodologies, nor is it to provide a full account of the feasibility or applicability of VSD to technology design. There has been much said in the existing body of literature that has already aimed to discuss those topics. Instead, this paper seeks to use the VSD methodology, particularly its emphasis on stakeholder values to provide a more general discussion of stakeholder involvement in technological design methodologies.

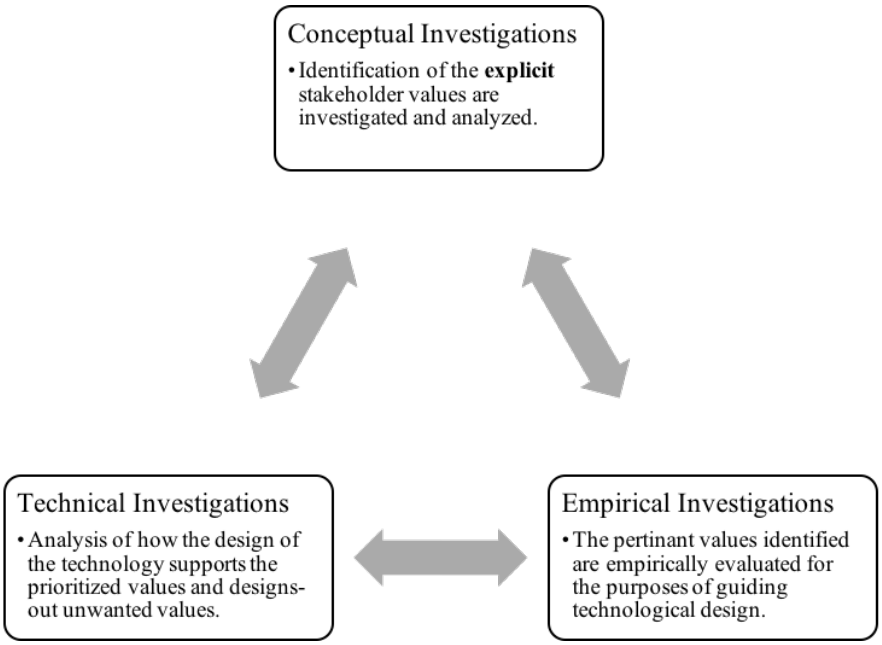

Figure 1. The tripartite methodology of VSD is self-relfexive and recursively self-improving

4. The moral psychology of conceptual investigations

What are values? Where do values come from? Which values are socio-culturally unique and which values universal? Which values can be integrated into the design of technological innovations? How do we balance apparently conflicting values such as autonomy and security? Should moral values always be given precedence over values that are non-moral? These are some of the issues that are addressed by the VSD approach, specifically by procedures outlined under conceptual investigations, one of the three investigations that compose the tripartite methodology of VSD (Friedman & Kahn Jr., 2002).

Conceptual investigations usually involve designers determining how both direct and indirect stakeholders might be affected by the technological innovation that they aim to design. Initial conceptual investigations may take the form of drawing upon the relevant literature of technologies that are applicably similar or involved in the technology that designers seek to develop. In the case of developing a nanopharmaceutical drug like a Doctor in a Cell, values such as privacy (medical data collected by the device), informed consent, safety and efficacy can be conceptualized as posing ethical and societal issues (Casci, 2004; Lucivero, 2012; Timmermans et al., 2011). The literature that analyses these particular issues, whether it is the issues per se or in the context of some other technological innovation, can then be levied in order to understand better what the impacts of those issues may be in relation to nanopharmacy. The next step would be to use the initial results of the conceptual analysis to begin the technical work in designing the system. However, in order to successfully and holistically manage this, the epistemic status of the investigated values needs to be clearly demarcated as well as means by which that status can be buttressed by drawing upon heuristic tools.

4.1. The epistemology of conceptual analyses

The epistemic status of the values gathered by designers during conceptual investigations seems, as a result of the theories proposed by moral psychology, dubious. If we assume that value judgments are in fact the result of a cognitive process and that reasoning is an ex post facto activity, then the susceptibility of cognition to various cognitive biases makes the resulting value judgments less credible. Yet, the value judgments are a critical part of the VSD methodology. How can we reconcile the apparent lack of value-grounding in the VSD approach?

Van Wynsberghe's (2013) approach to this realization was to conceptualize the origin of value in the VSD framework. She attempted to add a level of normativity to the ethical valuations that are essential to the conceptual investigations of VSD. Her application of VSD to the design of care robots was thus argued to be grounded on the existent values of care. Two separate arguments can be made regarding this manoeuvre: 1) the VSD methodology is meant to be augmented in ways that best integrate it with current practices and activities, thus her augmentation was simply a spirited instantiation of VSD philosophy, and 2) VSD methodology is grounded in normative ethics by taking as its foundation "moral value such as freedom, equality, trust, autonomy or privacy justice [that] is facilitated or constrained by technology" (Friedman, 1997; van den Hoven, 2013, 137). Hence, Van Wynsberghe's adoption of values specific to care could be viewed simply as field-specific interpretations of the already grounding values that VSD holds as foundational, thus making her move nothing more than a context-sensitive focusing of existing values.

The spirit of this approach, however, opens up what could ultimately prove to be a detrimental stance against adopting a VSD methodology. Rather than designers having to manipulate the VSD approach for every particular technological innovation that they are attempting to develop responsibly, I propose that the solution to the susceptibility of moral judgments on account of cognitive biases does not lie in the search for a normative moral foundation but rather can be solved by adding a new methodological tool to the stage of conceptual analyses. Cognitive biases influence our moral intuitions, particularly in relation to controversial technologies (Caviola et al., 2014). Thus, to more authentically ascertain the moral intuitions of individuals, and in turn gain a better grasp of the intersubjective values that they hold regarding a particular technology, it is a useful practice for designers, during their conceptual investigations, to employ certain psychological heuristic practices that reduce the influence of cognitive biases toward technology (see Bostrom & Ord, 2006; Larrick, 2004; Savulescu, 2007).

Hence, the goal of grounding VSD in objective/universal values is not only dubious, given what I have discussed regarding intuitionism, but unnecessary. The current need for a DFV, like VSD, for transformative technology creates a time-sensitive imperative to instantiate a design methodology *before* the technologies are developed, many of which are currently in development. As such, a pragmatic impetus exists to regarding the existent DFV approaches, those being: the realization of the intersubjectivity of moral intuitions, the lack of a need for value objectivity and the need to de-bias stakeholder value intuitions. Heuristic tools can help designers with the latter.

4.2. A heuristic toolkit

Given that conceptual investigations in VSD require an analysis of the ethical literature available to better understand the moral and non-moral values of stakeholders at play,

a good starting point for the designers who seek to apply a VSD methodology would be to acknowledge the theoretical underpinnings of the moral epistemology of the values investigated. This means that developers understand how moral judgments are made and that cognitive biases affect moral judgments. In light of this, conceptual analysis should not only account for the ethical literature at play, but also the psychological literature and relevant scientific evidence that can be levied to better justify which values are included in the design as well as how tradeoff values are balanced (Caviola et al., 2014). Likewise, remedial measures must also be put into play in order to better judge which values are most authentic and also to create impartial evaluations through employing simple heuristic tests.

One such heuristic test is Bostrom and Ord's (2006) Double Reversal Test that aims at reducing the status quo bias in its judgments regarding technological innovation. They describe the effectiveness of the Double Reversal Test in its applicability to cognitive enhancement technologies saying that:

The Double Reversal Test works by combining two possible perceptions of the status quo. On the one hand, the status quo can be thought of as defined by the current (average) value of the parameter in question. To preserve this status quo, we intervene to offset the decrease in cognitive ability that would result from exposure to the hazardous chemical. On the other hand, the status quo can also be thought of as the default state of affairs that results if we do not intervene. To preserve this status quo, we abstain from reversing the original cognitive enhancement when the damaging effects of the poisoning are about to wear off. By contrasting these two perceptions of the status quo, we can pin down the influence that status quo bias exerts on our intuitions about the expected benefit of modifying the parameter in our actual situation. (Bostrom & Ord, 2006, p. 673)

Hence, its purpose, as Bostrom and Ord clearly state, is to attempt to determine exactly how the status quo bias influences intuition. In doing so, we can better understand exactly how and why individuals argue for specific values. Designers whose aim it is to apply the VSD methodology as thoroughly as possible need to approach their conceptual investigations with the additional activity of de-biasing their moral valuations. Because transformative technologies are more likely to elicit moral intuitions that have a higher likelihood of being influenced by biases, as a consequence of the controversial nature of the technologies (Caviola et al., 2014).

As such, what is required of designers is a holistic account of how to responsibly innovate through a DFV methodology. The VSD framework provides a sound basis from which to start, however fundamental characteristics of the method that may, from the outset, be susceptible to criticism need to be addressed before we can confidently adopt the approach ubiquitously. Work is already being done on the status of moral intuitions in making judgments about technology (e.g., Klein, 2016), as has been the role that cognitive biases play when making intuitive judgments about transformative technologies (e.g., Caviola et al., 2014; Oliveira, 2009; Partridge, Lucke, Finnoff, & Hall, 2011). The VSD approach needs to begin by taking this literature into account and integrating it into the basic methodology rather than relying on designers to change the method in an ad hoc fashion for every potential application (although a basic element of change will always be necessary given the diverse range of applications). Addressing the issue of the contentiousness of moral intuitions not only strengthens the VSD and other stakeholder-centered approaches by reinforcing their moral grounding, but also to better understand the authentic values of stakeholders beyond the veil of bias interference.

There are a host of potential de-biasing methods that can be employed by designers,

each of which may be more applicable to a particular application than others. Because DFV approaches are principled and formulaic in their procedures, it is beyond this paper's scope in determining which tools are best suited to which application. As such, future research projects, such as the ones described in the proceeding section, should explore which methods are best employed in particular developmental streams.

5. Implications and Further Research

The main purpose of this paper was to show how DFV approaches, particularly those that centralize the position of stakeholders in the design of new technologies face an epistemological gape in determining the values that stakeholders express; that is, that the moral values that are of critical import to VSD (and values in general for DFV approaches) are subject to cognitive biases. Because one of the primary means by which stakeholder values are extracted is by simply asking stakeholders what they value, the lack of a priori moral reasoning that moral psychology shows possess an issue for arriving at a more authentically informed RI.

These novel technologies, such as nanotechnology, biotechnology and artificial intelligence have been predicted to, at the very least, have major economic and societal impacts, unchecked development that lack explicit value engagement may prove catastrophic. As this paper opened, technology is inherently value-laden, and VSD takes this as its founding precept. Designing without values is impossible: whether or not they are deliberate is a matter of particular importance. As such, there is an urgency to direct the development of these transformative innovations towards beneficial futures viz. the embedding of pertinent stakeholder values and designing out those that run contrary to those beneficial futures.

This aim of this paper has been a humble one. Rather than offer a transformative or novel design methodology that seeks to encompass all of the values and issues that exist or may emerge, it has opted to offer a critique of current DFV methodology as they pertain particularly to the enrollment of stakeholder values as well as one of a potential number of ways to strengthen said methodologies. The use of heuristics in order to achieve a greater degree of authenticity regarding stakeholder values is but a simple, ad hoc, functional step that can be taken now. Further research projects should look at the viability of moral imagination theories that may be useful in bolstering the value-based investigations that DFV methodologies employ (see Boenink, Swierstra, & Stemerding, 2010; Lucivero et al., 2011; Mahoney & Litz, 2000). Doing so may be fruitful both prior and during the employment of DFV approaches. Because VSD aims to be self-reflexive and recursively improving, like many transformative technologies that can be directed through its use, its continually improvement, and even foundations restructuring through new research projects feeds into its modus operandi.

Additionally, further research into DFV approaches should look into their potential to anticipate not only emerging future values, but to also anticipate potential governance needs. As such the enrollment of stakeholders and the resultant value-integration may lead to novel and emerging governance structures and institutions. The potential for DFV approaches to anticipate potentially necessary governance mechanisms may be particularly salient as designers and other enrolled actors have a privileged position to inform policy makers of possible governance needs.

Finally, the application of de-biasing heuristics does meet particular constraints and contentions. The primary being their ability to be self-applied, that is, for designers and direct stakeholders who participate directly within the development and design of

technologies to self-apply these tools. Arguments could be made against this paper's thesis that there is a lack of symmetry in the operationalization of heuristics. This contention is methodological in nature, and requires a reformulation of the principles of DFV approaches in order to account for a symmetrical distribution and application of heuristic tools. As such, future research projects could explore how designers and developers can self-apply de-biasing tools in order to ensure that implicit and unwanted values are not designed into technologies. As such, the *designing out* of unwanted values plays a critical part of this as already mentioned, and because this is a principle of many DFV approaches – most explicitly VSD – tools to ensure its success are of methodological importance.

6. Conclusion

Although further research on these issues may show that we do need new normative frameworks for emerging technologies, as things currently stand it is unclear what those moral frameworks could or should be. Instead, we should focus on the very real and pressing issues that exist given that transformative technologies are already heavily funded, their development underway, and their convergence already being experienced. In light of the pragmatic imperative that now exists, it is up to us to determine how we can intervene in the development of these transformative, emerging and converging technologies to direct them in such a way that is aligned with the values of stakeholders. The VSD approach is one such methodology that aims to incorporate the values of stakeholders during the early design phases. However, the current VSD methodology, although accepting of augmentation, requires far too much ad hoc manipulation for integration into existing design and development practices.

What is needed is a reimagining of DFV frameworks. This reimaging, regarding VSD in this case, preserves the tripartite investigations of the approach while adding a critical tool to conceptual investigations. By adding heuristic tools to the conceptual analyses of values, the VSD methodology is strengthened against doubts about the epistemic status of moral judgments produced by moral intuitions. Although this paper has not shown that the employment of heuristics creates epistemic certainty regarding moral judgments, it has demonstrated that in light of the pragmatic urgency, as well as the current status of the origin of moral judgments, the implementation of practices that aim to de-bias moral judgments is critical to the success of responsible innovation viz. DFV approaches.

7. References

Aad Correlje; Eefje Cuppen, Marloes Dignum, U. P. & B. T. (2015). Responsible Innovation in Energy Projects: Values in the Design of Technologies, Institutions and Stakeholder Interactions 1 (Draft version for forthcoming book) Aad Correljé, Eefje Cuppen, Marloes Dignum, Udo Pesch & Behnam Taebi. In B.-J. Koops, I. Oosterlaken, H. Romijn, T. Swierstra, & J. van den Hoven (Eds.), *Responsible Innovation 2* (pp. 183–200). Springer International Publishing. Retrieved from https://link.springer.com/chapter/10.1007%2F978-3-319-17308-5_10

Alvial-Palavicino, C. (2016). The Future as Practice. A Framework to Understand Anticipation in Science and Technology. *TECNOSCIENZA: Italian Journal of Science & Technology Studies*, *6*(2), 135–172. Retrieved from

http://www.tecnoscienza.net/index.php/tsj/article/view/239

Boenink, M., Swierstra, T., & Stemerding, D. (2010). Anticipating the Interaction between Technology and Morality: A Scenario Study of Experimenting with Humans in Bionanotechnology. *Studies in Ethics, Law, and Technology*, *4*(June 2016), 1. https://doi.org/10.2202/1941-6008.1098

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press. Retrieved from https://global.oup.com/academic/product/superintelligence-9780199678112?cc=ca&lang=en&

Bostrom, N., & Ord, T. (2006). The reversal test: eliminating status quo bias in applied ethics. *Ethics*, *116*(4), 656–79. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17039628

Brey, P. A. E. (2012). Anticipatory Ethics for Emerging Technologies. *NanoEthics*, *6*(1), 1–13. https://doi.org/10.1007/s11569-012-0141-7

Briggs, P., & Thomas, L. (2015). An Inclusive, Value Sensitive Design Perspective on Future Identity Technologies. *ACM Transactions on Computer-Human Interaction*, *22*(5), 1–28. https://doi.org/10.1145/2778972

Brink, D. O. (2014). Principles and Intuitions in Ethics: Historical and Contemporary Perspectives. *Ethics*, *124*(4), 665–694. https://doi.org/10.1086/675878

Brown, S. (2009). The new deficit model. *Nature Nanotechnology*, *4*(10), 609–611.

Capurro, G., Longstaff, H., Hanney, P., & Secko, D. M. (2015). Responsible innovation: an approach for extracting public values concerning advanced biofuels. *Journal of Responsible Innovation*, *2*(3), 246–265. https://doi.org/10.1080/23299460.2015.1091252

Casci, T. (2004). Doctor in a cell. *Nature Reviews. Genetics*, *5*(6), 406.

Cash, D. W., Clark, W. C., Alcock, F., Dickson, N. M., Eckley, N., Guston, D. H., ... Mitchell, R. B. (2003). Knowledge systems for sustainable development. *Proceedings of the National Academy of Sciences*, *100*(14), 8086–8091.

Caviola, L., Mannino, A., Savulescu, J., & Faulmuller, N. (2014). Cognitive biases can affect moral intuitions about cognitive enhancement. *Frontiers in Systems Neuroscience*, *8*(October), 1–5. https://doi.org/10.3389/fnsys.2014.00195

Chilvers, J. (2007). Deliberating Competence: Theoretical and Practitioner Perspectives on Effective Participatory Appraisal Practice. *Science, Technology, & Human Values*, *33*(2), 155–185. https://doi.org/10.1177/0162243907307594

Collingridge, D. (1980). *The social control of technology*. Frances Pinter. Retrieved from https://books.google.ca/books/about/The_social_control_of_technology.html?id=2q_uAAAAMA

Cosmides, L., & Tooby, J. (1992). Cognitive Adaptations for Social Exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the generation of Culture* (pp. 163–228). New York: Oxford University Press. Retrieved from http://www.cep.ucsb.edu/papers/Cogadapt.pdf

Cummings, M. L. (2006). Integrating ethics in design through the value-sensitive design approach. *Science and Engineering Ethics*, *12*(4), 701–715. https://doi.org/10.1007/s11948-006-0065-0

Cushman, F., Young, L., & Hauser, M. (2006). The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm. *Psychological Science*, *17*(12), 1082–1089. https://doi.org/10.1111/j.1467-9280.2006.01834.x

Daniel, S. (2002). Real-Time Technology Assessment. *Technology in Society*, *24*, 93.

Dechesne, F., Warnier, M., & van den Hoven, J. (2013). Ethical requirements for reconfigurable sensor technology: a challenge for value sensitive design. *Ethics and Information Technology*, *15*(3), 173–181. https://doi.org/10.1007/s10676-013-9326-1

Delgado, A., Kjølberg, K. L., & Wickson, F. (2010). Public engagement coming of

age: From theory to practice in STS encounters with nanotechnology. *Public Understanding of Science*, *20*(6), 826–845. https://doi.org/10.1177/0963662510363054

Foley, R. W., Bernstein, M. J., & Wiek, A. (2016). Towards an alignment of activities, aspirations and stakeholders for responsible innovation. *Journal of Responsible Innovation*, *3*(3), 209–232. https://doi.org/10.1080/23299460.2016.1257380

Friedman, B. (1997). *Human Values and the Design of Computer Technology*. (B. Friedman, Ed.). CSLI Publications. Retrieved from https://web.stanford.edu/group/cslipublications/cslipublications/site/1575860805.shtml#

Friedman, B., & Kahn Jr., P. H. (2000). New Directions: A Value-sensitive Design Approach to Augmented Reality. In *Proceedings of DARE 2000 on Designing Augmented Reality Environments* (pp. 163–164). New York, NY, USA: ACM. https://doi.org/10.1145/354666.354694

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., & 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

Greene, J. D. (2014). Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics. *Ethics*, *124*(4), 695–726. https://doi.org/10.1086/675875

Guston, D. H. (2014). Understanding "anticipatory governance." *Social Studies of Science*, *44*(2), 218–242. https://doi.org/10.1177/0306312713508669

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037//0033-295X

Huldtgren, A. (2014). Design for Values in ICT. 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. 1–24). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6994-6_35-1

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2013). Motivated Numeracy and Enlightened Self-Government. *Yale Law School, Public Law Working Paper*, (307), 54–86. https://doi.org/10.2139/ssrn.2319992

Kaiser, M., Kurath, M., Maasen, S., & Rehmann-Sutter, C. (2009). *Governing future technologies: nanotechnology and the rise of an assessment regime* (Vol. 27). Springer Science & Business Media.

King, M., Whitaker, M., & Jones, G. (2011). Speculative Ethics : Valid Enterprise or Tragic Cul-De-Sac ? In A. Rudnick (Ed.), *Bioethics in the 21st Century* (pp. 139–158). InTech. https://doi.org/10.5772/19684

Klein, W. E. J. (2016). Problems with moral intuitions regarding technologies. *IEEE Potentials*, *35*(5), 40–42. https://doi.org/10.1109/MPOT.2016.2569742

Larrick, R. P. (2004). Debiasing. In *Blackwell Handbook of Judgment and Decision Making* (pp. 316–338). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470752937.ch16

Lucivero, F. (2012). *Too good to be true? Appraising expectations for ethical technology assessment*. (P. A. E. Brey, M. Boenink, T. E. Swierstra, & M. Boenink, Eds.). Enschede: Universiteit Twente. https://doi.org/10.3990/1.9789036533898

Lucivero, F., Swierstra, T., & Boenink, M. (2011). Assessing Expectations: To-

wards a Toolbox for an Ethics of Emerging Technologies. *NanoEthics*, *5*(2), 129–141. https://doi.org/10.1007/s11569-011-0119-x

Magnani, L. (2013). Abducing personal data, destroying privacy: Diagnosing profiles through artefactual mediators. In M. Hildebrandt & K. de Vries (Eds.), *Privacy Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology* (pp. 67–90). Routledge. https://doi.org/10.4324/9780203427644

Mahoney, J. T., & Litz, R. (2000). Moral Imagination and Management Decision Making. *Academy of Management Review*, *25*(1), 256–259. https://doi.org/10.5465/AMR.2000.2791616

Micheletti, C., & Benetti, F. (2016). Safe-by-Design nanotechnology for safer cultural heritage restoration. Retrieved December 15, 2017, from http://atlasofscience.org/safe-by-design-nanotechnology-for-safer-cultural-heritage-restoration/

Nichols, S., & Knobe, J. (2007). Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Nous*, *41*(4), 663–685. https://doi.org/10.1111/j.1468-0068.2007.00666.x

Nordmann, A. (2014). Responsible innovation, the art and craft of anticipation. *Journal of Responsible Innovation*, *1*(1), 87–98. https://doi.org/10.1080/23299460.2014.882064

Oliveira, J. R. (2009). Much ado about cognitive enhancement. *Nature*. https://doi.org/10.1038/457532b

Owen, R., Macnaghten, P., & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy*, *39*(6), 751–760.

Owen, R., Stilgoe, J., Macnaghten, P., Gorman, M., Fisher, E., & Guston, D. (2013). *A framework for responsible innovation.* (R. Owen, J. Bessant, & M. Heintz, Eds.), *Responsible innovation: managing the responsible emergence of science and innovation in society*. Wiley Chichester, Sussex.

Partridge, B., Lucke, J., Finnoff, J., & Hall, W. (2011). Begging Important Questions About Cognitive Enhancement, Again. *American Journal of Bioethics*, *11*(1), 14–15. https://doi.org/10.1080/15265161.2010.534536

Pinch, T., & Bijker, W. E. (1987). The social construction of facts and artifacts. In W. E. Bijker, T. P. Hughes, & T. Pinch (Eds.), *The Social construction of technological systems: new directions in the sociology and history of technology* (p. 405). MIT Press. Retrieved from https://books.google.ca/books?id=B_Tas3u48f8C&printsec=frontcover&dq=The+Social+Construction+Social Construction of Technological Systems&f=false

Rip, A. (2009). Technology as prospective ontology. *Synthese*, *168*(3), 405–422. https://doi.org/10.1007/s11229-008-9449-9

Rip, A., & Van Amerom, M. (2009). Emerging de facto agendas surrounding nanotechnology: two cases full of contingencies, lock-outs, and lock-ins. In *Governing future technologies* (pp. 131–155). Springer.

Roache, R. (2008). Ethics, speculation, and values. *NanoEthics*, *2*(3), 317–327. https://doi.org/10.1007/s11569-008-0050-y

Sarewitz, D., & Pielke, R. A. (2007). The neglected heart of science policy: reconciling supply of and demand for science. *Environmental Science & Policy*, *10*(1), 5–16.

Savulescu, J. (2007). Genetic Interventions and The Ethics of Enhancement of Human Beings. In B. Steinbock (Ed.), *The Oxford Handbook of Bioethics*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199562411.003.0023

Shweder, R. A., & Haidt, J. (1993). The Future of Moral Psychology:

Truth, Intuition, and the Pluralist Way. *Psychological Science*, *4*(6), 360–365. https://doi.org/10.1111/j.1467-9280.1993.tb00582.x

Star, S. L. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist*, *43*(3), 377–391. https://doi.org/10.1177/00027649921955326

Taebi, B., Correljé, A., Cuppen, E., Dignum, M., & Pesch, U. (2014). Responsible innovation as an endorsement of public values: the need for interdisciplinary research. *Journal of Responsible Innovation*, *1*(1), 118–124. https://doi.org/10.1080/23299460.2014.882072

te Kulve, H., & Rip, A. (2011). Constructing Productive Engagement: Pre-engagement Tools for Emerging Technologies. *Science and Engineering Ethics*, *17*(4), 699–714. https://doi.org/10.1007/s11948-011-9304-0

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

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

van den Hoven, J. (2007). ICT and Value Sensitive Design. In P. Goujon, S. Lavelle, P. Duquenoy, K. Kimppa, & V. Laurent (Eds.), *The Information Society: Innovation, Legitimacy, Ethics and Democracy In honor of Professor Jacques Berleur s.j.: Proceedings of the Conference "Information Society: Governance, Ethics and Social Consequences", University of Namur, Belgium 22–23 May 20* (pp. 67–72). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-72381-5_8

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. Retrieved from https://books.google.ca/books?id=VVM_AAAAQBAJ&dq=moral+value+such+as+freedom,+equality,+

van den Hoven, J. (2014). Nanotechnology and Privacy: The Instructive Case of RFID. In R. L. Sandler (Ed.), *Ethics and Emerging Technologies* (pp. 285–299). London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137349088_19

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., & Weckert, J. (2008). *Information Technology and Moral Philosophy.* (J. van den Hoven & J. Weckert, Eds.). Cambridge University Press. Retrieved from http://www.cambridge.org/catalogue/catalogue.asp?isbn=9780521855495

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. (2016). Service robots, care ethics, and design. *Ethics and Information Technology*, *18*(4), 311–321. https://doi.org/10.1007/s10676-016-9409-x

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

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a Bomb on a Person Versus Throwing a Person on a Bomb: Intervention Myopia in Moral Intuitions. *Psychological Science*, *18*(3), 247–253. https://doi.org/10.1111/j.1467-9280.2007.01884.x

Winner, L. (2003). Do artifacts have politics? *Technology and the Future*, *109*(1), 148–164. https://doi.org/10.2307/20024652

Woodward, J., & Allman, J. (2007). Moral intuition: Its neural substrates and normative significance. *Journal of Physiology - Paris*, *101*(4), 179–202.

https://doi.org/10.1016/j.jphysparis.2007.12.003

Acknowledgements

I would like to thank Lorenzo Magnani for providing feedback on an earlier draft and the two anonymous reviewers for their constructive comments that helped to produce a more rigorous manuscript. Any errors are the authors' alone. The views in the paper are the authors' alone and not the views of the Institute for Ethics and Emerging Technologies.

Author biography

Steven Umbrello Steven Umbrello is the Managing Director of the Institute for Ethics and Emerging Technologies and a researcher at the Global Catastrophic Risk Institute with research interests in explorative nanophilosophy, the design psychology of emerging technologies and the general philosophy of science and technology