A Value-Sensitive Design Approach to Intelligent Agents

A Value-Sensitive Design Approach to Intelligent Agents

Steven Umbrello[a] and Angelo Frank De Bellis[b]

[a] Institute for Ethics and Emerging Technologies; [b] University of Edinburgh

ARTICLE HISTORY
Forthcoming chapter in *Artificial Intelligence Safety and Security* (2018) CRC Press (.ed) Roman Yampolskiy.

ABSTRACT
This chapter proposed a novel design methodology called Value-Sensitive Design and its potential application to the field of artificial intelligence research and design. It discusses the imperatives in adopting a design philosophy that embeds values into the design of artificial agents at the early stages of AI development. Because of the high risk stakes in the unmitigated design of artificial agents, this chapter proposes that even though VSD may turn out to be a less-than-optimal design methodology, it currently provides a framework that has the potential to embed stakeholder values and incorporate current design methods. The reader should begin to take away the importance of a proactive design approach to intelligent agents.

KEYWORDS
artificial agents; artificial intelligence; value sensitive design; design psychology; ethics

1. Introduction

The field of artificial intelligence development is experiencing a productive period of incredible progress and innovation, and any consequences, whether positive or otherwise, will fall on the designers' shoulders. As such, the onus is on us to shape technological advancement in such a way as to mitigate any foreseeable and, as much as possible, unforeseeable catastrophic events. After several boom and bust periods, AI is now primed to be one of our most critical advancements as well as one of our most dangerous ones. The impending emergence and proliferation of AI into the public sphere will assuredly result in an encompassing revolution of sorts—one that will affect society and the ways in which both humans and nonhumans work and communicate, one that promises to change our quotidian lives by relieving us of menial tasks, and one that will fundamentally change how humans perceive ourselves, others, and the world – and vice versa.

AI designers have already begun in earnest to develop sophisticated systems that mimic, and in certain cases surpass, human intelligence: such advancements are capable of learning, interacting with their surroundings, and making novel decisions, all in autonomous manners and when faced with highly variable situations. Because of this level of intelligence, the agents do and will possess—their increasing capacity to receive information about their environment, decide, and act—the obvious question arises: Who is responsible if something goes wrong? But perhaps the question that

Corresponding author: steve@ieet.org (Steven Umbrello)

should be asked is, how can technology be made to advance such that it is encouraged to grow and simultaneously herded to progress in a manner that doesn't lead to disaster? After all, responsibility is often an afterthought; it typically begins after something has gone awry, when intentional guidance could have reduced or eliminated unwanted consequences along with the clumsiness of deferring disaster to an individual or homogenized entity.

In their ability to reason and behave using human-like logic, these intelligent agents become actors in the real world where they may interact with human agents and even engage in activities they are given complete control over. Self-driving vehicles, military robotics, smart home devices, big data analysis software, and care robots are just a few examples of technologies that already use artificially intelligent processes, albeit many are currently rudimentary, to afford us the opportunity to rely on robotic devices to complete tasks for users that they can't do on their own, or to afford the luxury of not having to engage in the completion of routine duties.

Current trends in the development of intelligent agents illustrate not only the extent to which existing systems are autonomous but also how increased autonomy will continue to decrease the amount of human supervision necessary to their normal functioning. This abdication of certain activities along with the human-like qualities AI will bring should make us feel a sense of responsibility; the progression of the technology should be a call to arms to ensure that due consideration is given to their development. This becomes truer as the range of abilities of intelligent agents widens. As the breadth of functional capacity grows, it becomes imperative that designers take steps to design systems that will not act dangerously or undermine stakeholder interests. Even if international command and control mechanisms and regulations are put into practice to ensure the safe and consistent behavior of agent, multi-agent, and human-agent level systems, ethical issues associated with their design and implementation are still likely to remain complex, with additional emerging issues sure to arise that will implicate a broad range of values, moral issues, and ethical principles.

Of course, further complexities spawn from the implementation of ethical approaches taken when designing or constructing physical entities or AI. The choosing of ethical frameworks and the intricacies of value alignment become all the more pressing as intelligent agents confront and are exposed to novel environments, grow in human-nonhuman society, interact with other agents based on different design principles, act on behalf of people, and share common resources (Taylor et al. 2016; Soares and Fallenstein 2014; see also van den Hoven and Jacob 2013).

To tackle these issues, this chapter proposes that a particularly suitable design approach to combat these complexities while continuing to permit the steadfast nature of scientific development is one that allows involved parties to visualize and take into consideration important human values such as safety, privacy, accountability, and sustainability. In addition to these considerations, the design approach must also—and this is what separates it from rigid value approaches or responsibility attribution methods—beg for involved parties to prepare for moral conflicts as and before the AI development is underway. Moreover, the value-laden design method should be flexible in that it can swiftly adapt to value trade-offs and moral overload (Van den Hoven, Lokhorst, and Van de Poel 2012). For instance, when designing a self-driving vehicle, designers may choose to impart it with human-like driving behaviors instead of the behaviors one would anticipate—a vehicle with strict, precise measures to drive in a calculated manner based on road rules. Though initial planning stages may have designers agree that safety is of utmost importance, real-world experimentation may yield results that argue in favor of self-driving vehicles responding to and behaving as if a real human

driver were manning it. Accomplishing this may ultimately require balancing deontic rule-following, utility maximization, and risk assessment in the agent's logic to achieve the ultimate goal of roadway safety. Thus flexibility and prevalence during design and execution make for critical parameters of a design methodology, one that applies to AI.

Given these considerations, the design methodology that most suits what we believe to be an all-encompassing, flexible, and capable one is known as value-sensitive design (VSD). Though various design methodologies could be employed in order to safely design intelligent agents—participatory design, universal design, user-centered design, and inclusive design—the aim of this chapter is to introduce VSD as a potential and undeniable candidate for carefully and unrestrainedly approaching the design of beneficial intelligent agents. Although future research projects may show that another design approach or an amalgamation of existing methods may be more suitable, this chapter aims to show how contemporary design psychology and methodology may provide the most appropriate starting framework, given the speed of development and the risks associated with unmitigated engineering (Baum 2016; Muehlhauser and Bostrom 2014; Soares and Fallenstein 2014).

2. An Introduction to Value-Sensitive Design

Before applying the design theory and proving how it can be used to the benefit of ethical and responsible methods of producing intelligent agents, it is important to have an understanding of the theory itself, its main components, and its limitations.

The most notable literature written on VSD has been produced by Batya Friedman, in which she comprehensively breaks down the design methodology to both illustrate what it is and how it can be applied to fields dealing with digital technologies, technologies that affect humans physiologically and psychologically, and intricate prediction software that influence the future of our world. She sums up VSD 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., and Borning 2006, 1; see also Friedman 1996; Friedman and Kahn 2002). Probably the most notable aspect of this definition and one that was previously detailed is the fact that VSD is a design methodology that does not take place during any one instance during the design phases—it occurs during all of them.

Prior to Friedman's outlining of VSD and its applications, there have been scarce amounts of context given to such design methods, their theories, and implementations. She lays out a particularly useful set of methods, which she calls an integrative and iterative tripartite methodology, for framing the three non-sequential activities essential to VSD: conceptual investigations, empirical investigations, and technical investigations. It is important to note that these three investigations combine to form the cycle of duties involved in the appropriate application of VSD to the field of interest.

The theoretical approach of VSD and its tripartite investigations set it apart from other design approaches because it is self-reflexive, fallible, and continually improving. Unlike the typical methods of ascribing responsibilities to certain parties or groups of people, which often only account for issues after they arise, VSD takes a more proactive approach. That is, its value-focus is not singular or temporally locked, but it seeks to predict emerging values and issues and in response "influences the design of technology early in and throughout the design process" (Friedman, Kahn Jr., and Borning 2006, 12). Such approaches break down the fear caused by merit-based perspectives to re-

sponsibility and the potential scenarios of bribery found in ascribing responsibility in rights-based perspectives (Doorn 2012).

Merit-based approaches, as Doorn (2012) argues, take a retributivist angle on how and who should be held liable. The person or people involved in the misdemeanor or success of a technology, sentient or otherwise, ultimately receive the praise or the blame in merit-based approaches. This is a dangerous, narrow-minded method of delegating responsibility, and for two reasons: (1) It instills fear in those creating new technologies to explore creative, risky opportunities and (2) there exists categorical differences in the way in which responsibility can be broken down—moral responsibility vs. causal responsibility. It is the latter of the two reasons, with its critical distinction, that will be discussed in further detail in below.

Causal responsibility could be anything from sheer luck to some negative happenstance that a person has no reasonable control over, whereas moral responsibility has more to do with those involved in the creation process of a particular technology. Ascribing responsibility to either of the two, moral and causal responsibility, becomes muddied when it comes to technologies—and especially, in this case, intelligent agents—with several variables and a vast array of professionals providing input in the creation process. When something goes wrong, should the blame fall both on the user and the creator? And if so, how many of those parties involved in the creation should be held liable? A merit-based approach, like a rights-based approach, both hinge on methods focused on defining responsibility and blameworthiness.

In a rights-based approach, the focus is placed on the rights of the users, those that could be potential victims for adverse outcomes caused by the technology. Like a merit-based approach, this perspective pays close attention to ethical complications after one has been transgressed. It aims to be remedial. Actions are imperative in the rights-based approach, where how the end users are impacted matter greatly. The two driving forces behind the approach are strict liability (a legal requirement) and informed consent.

The first, strict liability, serves to identify which person was involved in the outcome that posed direct harm to a victim. Taking what has been explained about merit-based responsibility, it's quite clear that ascribing accountability in this way becomes tricky when the person involved may have triggered it by accident. Nonetheless, the rights-based approach identifies all parties, the creators, and the users, as those who carry the responsibility of ensuring that no harm is done. This is where the second driving force, informed consent, enters the discussion. Informed consent is a method of safeguarding certain parties from taking responsibility for certain outcomes. It advocates for the clear, unrestrictive explanation of possible outcomes to potential users so that the users themselves can decide whether they are comfortable with the benefits and possible hazards that follow the implementation of technologies.

Again, there are issues with this as well. While informed consent is a steady approach when it comes to forging a relationship between creators and users, as it calls for the comprehension and voluntary participation of users, it fails when it comes to making technological progress in a frictionless manner. To clarify, the final say is up to the users to decide whether to proceed with the advancement and doing so ultimately means that if most of a collective group agree that the technology is too risky, it won't be developed (Doorn 2012). While this certainly makes sense and is a sturdy safeguarding method to keep irresponsible, unethical design approaches and technologies at bay, it also opens the issue of bribery. In a move to appease potential users, those developing the new, potentially uncertain technologies may resort to providing monetary compensation to those members of the user group who have vetoed the development. Of course, the

other issues it causes relate to the fact that many users, no matter how imperative informed consent makes it, will not fully understand what the new technology entails today or in the years to come.

These responsibility ascriptions are important to understand because they highlight the issues involved when determining methods for protecting humans from unintended or unlucky consequences imposed by new technologies. In an effort to elucidate the effectiveness of VSD, we have detailed accounts of the merit-based and rights-based approaches responsibility because their shortcomings, though having some desirable outcomes, are avoided by its tripartite methods. Whereas the responsibility ascriptions focus mainly on future outcomes or sometimes hard stops in the initial phases, VSD works from both dimensions. The tripartite method that will be discussed in detail—conceptual investigations, empirical investigations, and technical investigations—collaborate to account for consequences as they arise during and after development.

Friedman explains several qualities about VSD that set it apart from other design methods, and by extension, we argue that these distinct qualities also contrast VSD concerning the traps set by merit-based and rights-based perspectives. These include classes like proactivity, diverse areas of application, widened scope of stakeholders (both direct stakeholders—such as designers, patients, and users—and indirect stakeholders, such as the general public and corporations), and careful attention to universally held values (Timmermans, Zhao, and van den Hoven 2011). There are more, but we think these are the most critical when it comes to grasping the stark differences and benefits posed by this design methodology.

Beginning with proactivity, this quality is the impetus of which all the tripartite methodologies are based on. Without proactive measures, advancements are met with the necessity to take correctional measures after the fact, which of course involves a merit-based approach. VSD avoids these issues of determining the difference between moral responsibility and causal responsibility. When it comes to a rights-based ap- proach, though one could argue that it too behaves proactively, the insecurity of such a method is that it focusses too much on a single aspect of the stakeholder and is more of an actions-based method than one drawn from an approach that is theoretically grounded and that takes more than one party into consideration.

The second differentiating quality deals with the extent to which VSD is more diverse than other design methods, and as we argue, the frameworks set by the two discussed responsibility attributions. Friedman explains, ". . . Value Sensitive Design enlarges the arena in which values arise to include not only the work place, but also educations, the home, commerce, online communities, and public life" (Friedman, Kahn Jr., and Borning 2006, 13). It is critical that it extends to all these areas, especially when it comes to the engineering and operationalization of values in design because such technologies naturally involve several and most often overlapping areas of concerned members. This also encroaches on the differences seen from VSD's widened scope of what it considers stakeholders. While other design approaches, which will not be discussed in detail here, employ either method of collaboration or democratic values, VSD tries to capture the needs of several groups comprehensively and doesn't exclude certain parties just because their needs are not in the majority, like the Inclusive Design methodology (see Briggs and Thomas 2015; Clarkson et al. 2003; Keates and Clarkson 2003; Newell et al. 2011).

Finally, VSD takes into account the importance of universally held values. This too is critical when dealing with AI technologies, as they will certainly be used across varying, diverse cultures and moral systems. Universally held values, as described by Friedman,

arise when discussing the seemingly different values held by various cultures, but that have similar core values (whether this is actually true is a worthwhile future research project). What one culture does to promote etiquette, another culture may express differently. VSD aims to account for this and requires that the development of technologies acknowledge values that appear different but that are nonetheless universally held. Although the philosophical grounding of universal values is highly contentious in the philosophical literature, the VSD methodology seeks to ground its ascription of universality on the common root values that are instantiated differently in various cultures (Friedman 1996; Friedman and Kahn Jr. 2002).

Equipped with the knowledge of what VSD is and what it does differently to overcome the shortfalls of certain responsibility ascriptions, it is critical to understand how its three design perspectives each add to the combined, all-encompassing nature of the approach. The three—conceptual, empirical, and technical— investigations lend themselves to the proactive nature of the design. That is, each of the three carries each other along, and when one area fails, it likely causes all other areas to require some taking care of as well. Each of the three cannot stand on its own, and all of them must be accounted for when taking a value-sensitive approach to design.

2.1. Conceptual Investigation

The first of the three, conceptual, is the more theoretical aspect of VSD. In short, conceptual investigations are those that are informed by philosophy. This area of engagement involves answering the critical questions related to defining certain parameters before and during technological developments, the 5 Ws: Who are the stakeholders? What are the related values in this scenario? Where do certain parameters begin and end when discussing the bounds of usability versus privacy/safety and direct versus indirect stakeholder? When are the agreed upon methods and procedures no longer viable or in support of the values being sought? Why is one design supported and another excluded? These questions and other theoretical areas are explored in conceptual investigations (see Denning, Kohno, and Levy 2013 for a good example of the VSD application to computer security and smart homes. This is particularly illustrative of intellegent systems).

Part of Friedman's exploration of conceptual investigations include a real-world application for cookies used in web browsers. In her discussions, she evaluates the importance of informed consent in such technologies so that users know what cookies are and what their function is, so that such members have full control over whether they accept the use of such tracking software. Though we have mentioned the dangers of informed consent when drawing on literature related to responsibility, this example provides the opportunity to further bolster the effectiveness of the tripartite methodology adopted by VSD. Because there are inherent shortcomings to permission architectures like informed consent, such as bribery or crippling fear, not only does VSD compensate by including other investigations, but informed consent on its own is certainly not the only discussion part of the conceptual third. Conceptual methods involve research, the background knowledge that other investigations will draw from, and while informed consent on its own may lead to hard stops or unfair blame, it combines with the remainder of the perspectives to correct and assess advancements from multiple areas of need.

2.2. Empirical Investigation

The second of the three investigations is empirical. As the name suggests, an empirical investigation involves taking account of quantifiable measures relating to the potential success or failure of a particular development. The variables quantified could be things like statistical data that describes patterns of human behavior, assessments that mea- sure the needs and wants of the users, and the dichotomy between what people say they want in a design and what they actually care about in practice (Friedman, Kahn Jr., and Borning 2006). An example of such an instance where users contend with not truly knowing what their expectations for a new technology are, can be illustrated by those who ask for privacy by having a prompt give them control over whether to allow cookies to be used in web browsers. When this prompt runs, some users may take the time to read it and make a conscious decision, while others may, in an effort to browse without disruption, skip past or ignore the message. Whereas conceptual investigations are more like background evidence and qualitative support for a technological design endeavor, empirical methods extend the research to precise measurements to prove, resolve, or progress design work. Of course, investigating the potential impact imposed by technologies through quantitative and qualitative means falls short if the materials themselves, that is, the technical nature of the devices isn't explored.

2.3. Technical Investigations

The final investigation is technical. Technical investigations involve considering the actual materials and the nature of the technology. For example, when developing intel- ligent agents, one would want to consider the delivery method: should the technology be embodied within a shell like a robot? Should it simply be code delivered to com- puting systems? Should it exist within several household items like other Internet of Things (IoT) devices? The technical questions become important in the operational- ization of values given that they can constrain how those values are instantiated in the design.

All of these kinds of questions form the basis of technical investigations because they ask about the nature of the technology and their abilities to support values. As Friedman says ". . . a given technology is more suitable for certain activities and more readily supports certain values while rendering other activities and values more difficult to realize" (Friedman, Kahn Jr., and Borning 2006).

3. Applying VSD to Intelligent Agents Using the Tripartite Methodology

3.1. Conceptual Investigations and the Moral Dilemma

Perhaps the most difficult question we need to answer when it comes to researching, creating, and implementing intelligent agents is related to what it means to be moral and how morality differs or parallels ethics. What does it mean to be moral? Is being moral the same as being ethical? And, of concern here, is morality and ethics the same for intelligent agents—is our anthropic approach to morality ethical or fair to endow on intelligent agents? Should they have their own class of what is right and what is wrong?

Keith Abney (2014) defines this problem well when he explains that when developing autonomous robotics—and it's similar for intelligent agents in general—there are three ways to define ethics: (1) the ethical practices of the human creators, users, owners,

customers; (2) the principles that are endowed to the robot; and (3) the transcendence of the robot to think for itself and therefore follow its own ethical framework. And to take his definition of morals versus ethics, morals are actions that "ought" to be followed or avoided, and ethics are the rules for and the study of moral beliefs. Conceptual investigations involve researching and understanding specific scenarios where technologies are to be used, and for as long as intelligent agents are not quite as advanced to make conscious decisions as a human would (i.e., a “deliberative system”), designers must consider that circumstance is an important consideration when making a choice to embed intelligent agents with moral programming.

Though the tripartite methodology can’t answer the question of how to create the perfect, here meaning moral, intelligent agent, it can aid in the development of the best possible, most comprehensive technological progress that behaves in a perceptibly ethical manner. Using the conceptual investigative component of VSD, research can be done to assess the stakeholders involved and their expectations of morality. Following conceptual investigations, aside from the correctional process of accounting for the other two areas of the tripartite combination, serves to lead engineers and ethicists away from following strictly their own beliefs of morality and instead inject morals that, at least on a base level, are universal (i.e., stakeholder sensitive), followed by ethical frameworks that embody the notions of morality within the specific area of use. It is this approach to design, which includes what Friedman calls values that are "universally held," that set it apart from other design methods (Friedman, Kahn Jr., and Borning 2006, 2). Conceptual design is but one of the facets of VSD, but it sets the stage for what is to follow by beginning with a concrete stance on what are deemed to be universally held values, which are then extended by methods to cater to circumstantial events.

3.2. The Need to Break Down Human Language

Though it is critical to construct a moral framework for intelligent agents when involved in the conceptual investigations, focusing solely on universally held values may, in fact, take away from the strengths of VSD’s deviation from rights-based approaches to responsibility. It is certainly critical to provide a rigid framework of dos and don’ts for intelligent agents, much like those given to young children, but restricting such creations to a finite set of conditionals creates a complicated mess of logic and breeds accusatory actions when something goes wrong (Soares 2016).

Though counteridenticals, as Bringsjord (2016) describes them, can provide a thorough understanding of intelligent agents through a series of grammatical frame- works—conditional sentences with subordinate clauses—the complexity of such con- structions increase with things like the involved verb mood. These can be used in intelligent agents so that they do what they ought to do, refrain from doing what they aren’t supposed to do, and sometimes do what is beyond expectation. While this sets up a critical construction for the creation of intelligent agents, it is also crucial that the design phases consider the import of circumstances that aren’t so well defined.

By relying merely on shallow or deep approaches like counteridenticals or counterfactuals, parties involved in creating the code to form the base of the agent’s decisions will surely miss some areas of uncertainty. Therefore, alone, counteridenticals are likely not to account for a vast range of scenarios like VSD approaches can. VSD, as will be discussed, plugs the holes left by qualitative research done in the conceptual investigation with empirical investigations.

3.3. A Controlled Approach

VSD's strength is not found in one individual method, but in the combination and the recursive nature of the three design investigations. Thus, it makes sense to introduce and add other methods of answering questions regarding morality. Perhaps looking to autonomous vehicles (AVs) is helpful when deciding on and providing for ethically sound intelligent agents.

Dogan et al. (2016) explain the importance of closing moral decision making so that decisions are made based on a limited amount of data. If designers consider human morality the goal when creating intelligent agents, then we would argue that people only deal with a limited amount of information at once when making decisions. There is also the fact that humans do make decisions rooted in emotion. Disregarding emotion, it is fair to say that creators of intelligent agents would likely want emotion, again depending on the circumstance, to be left out, although design frameworks like the independent core observer model (ICOM) may ultimately prove more beneficial in the long run (Waser 2016). This is where adding to the counteridentical framework described above becomes critical: there are too many variables to account for and too many ways of solving or behaving during a situation that limiting certain decisions seem to be the only way, particularly to mitigate any adverse outcomes that may result from a top-down hierarchical goal structure (Taylor et al. 2016b). That is unless the intelligent agent has transcended traditional artificial intelligence (Goertzel 2016; Rolf and Crook 2016).

Using VSD's conceptual approach, we can also account for what Dogan et al. (2016) define as the two clusters of values: self-transcendence and self-enhancement. He explains that in the creation of AVs, and this can arguably be extended for any autonomous type of technology, these two values are the driving force behind their behaveior. The first values the collective, so designers would endow intelligent agents with the goal of offering protection for the greater good, while the second cluster of values involves accounting for individual interests. Again, depending on the context, as studied in conceptual investigations, it is important to consider the technology being developed to decide if the collective or the individual is of moral import. The flexibility offered by VSD allows for such considerations to balance the variables: positive, negative, good, bad, better, worse, moral, and immoral.

AVs often employ the use of principle-based approaches to their functionalities. That is, they behave according to rigid principles, usually traffic laws. However, case-based approaches are important as well since cases will often be circumstantial when dealing with such technologies like automated vehicles. Once again, the strength of a VSD approach has the potential to accommodate for all these aspects and could in fact be used, as we suggest, to intermingle both approaches as explained above. Perhaps a mixture of hard and fast laws, as those found in traffic laws, and case-based approaches to governing certain behaviours are critical when creating specific technological devices, and therefore for intelligent agents (see Bonnemains, Saurel, and Tessier 2016; Laurent Orseau and Armstrong 2016; Mermet and Simon 2016).

3.4. Empirical Investigations and the Moral Dilemma

Friedman outlines empirical investigations as any "activity that can be observed, measured, or documented" (Friedman, Kahn Jr., and Borning 2006, 4). As conceptual investigations investigate more of the background information regarding the viability of an intelligent agent, empirical investigations gain their importance from taking a

closer look at exactly how users and other stakeholder make use of the technology in their daily lives, the consequences of such purposes, the potential drawbacks of certain features, and other opportunity costs. Because users frequently ask for features that they don't actually use, empirical investigations help to balance what is included and what is excluded from final prototypes in order to mitigate dangers or promote better user experience.

Given the complexity and sophistication of intelligent agents it is necessary to define how empirical investigations are to be used. If we consider intelligent agents as having similar rights and responsibilities as humans, which can be argued using Abney's explanations of what constitutes personhood as an example (see Abney 2014), then it may not make sense to empiricize every behavior of the intelligent agent as if it was a tool to be used for some task.

The very ability of VSD to accommodate for and remain plastic in several scenarios lends itself well to such a case. Designers can apply empirics in some way when designing an intelligent agent, though it may be different than for other technologies. Friedman exemplifies this by explicitly outlining empirical investigations as they are used for technologies such as web browser cookies, television-screen windows, and urban sim software (Friedman, Kahn Jr., and Borning 2006). As such, it is important to note that VSD is applicable to various technologies, and can be applied to intelligent agents as well.

Take for example intelligent agents that serve multiple purposes, designers could measure how users, perhaps in a psychological sense, interact with the agents. Do humans accept them as moral beings or disregard them as artificially intelligent software shelled by chunks of metal and plastic? Of course, such research also combines with the technical investigations we will examine shortly, because the form and look of the intelligent agent is perhaps as important as what its functions are (see Breazeal 2003; Bourke and Duffy 2003; Fussell et al. 2008; Tao et al. 2008).

In another scenario, designers may perform empirical investigations by measuring the other critical data such as the measurable benefits the intelligent agents provide to the domain they work in, the increase in quality of life of those who employ the systems, and the consistency of ethical or non-ethical behaviours as set out during the conceptual investigations. Empirics are critical because they essentially back up the more quantitative evidence and research performed in the conceptual investigations. The empirical evidence is to take that background research and present a case for the usage of the technology, in this case the intelligent agents, in real-world situations.

3.5. Calculating Good

One of the more important considerations when creating ethical frameworks for robots is that there is great difficulty in calculating actions, or inactions, as being moral. That, plus it is ". . . an impossible demand to calculate the utility of every alternative course of action" (Abney 2014). Not only do intelligent agents face the aforementioned frame problem of first-order logic when facing weighty decisions, but so does behavior frameworks such as cost-benefit analyses as solutions to utilitarianism (see also Laurent Orseau and Armstrong 2016; Soares and Levinstein 2017; Soares, Yudkowsky, and Armstrong 2015; Sotala 2016; Taylor et al. 2016). Such reductive methods of quantifying and assigning values to things like love, devotion, and honor, is scarcely a suitable method as is the embedding of a moral framework that ignores these important human values (see Stocker 1976; Wolf 1982).

As with counteridenticals, the installation of intelligent agents cannot be supported using just empirical methods to uphold their value and validate their reasons for being created. As mentioned in the conceptual investigations portion of this chapter, such methods are perhaps critical in setting up a basic framework, but like VSD, other areas of research and experimentation is required to comprehend the limits of individual approaches entirely. Quantifying good for example is almost impossible because one person's good may not necessarily fit with another's—also what is considered good for an intelligent agent may not necessarily be considered good for a human (Allen 1997; Stahl 2004; Wiltshire 2015). Again, it all depends on the approaches taken, as guided by conceptual investigations, and what the outcome is to achieve.

Nonetheless, one of the most complex aspects of designing intelligent agents is that everyone wants them to be, first and foremost, moral beings, yet human agents are far from moral when it comes to the consistency of their behaviours (Dodig Crnkovic and Çürüklü 2012; Hellström 2013; Johnson and Axinn 2013; Nadeau 2006; Torrance 2013; Scheutz and Malle 2014; Sotala 2016). Abney (2014) details the oft-discussed two features of being considered moral persons: an inner moral sense and an emotional inner life. That is, humans make decisions based on emotion and deliberative, reasonable thought processes. While the emotional decisions are not considered a necessity when it comes to making moral decisions, it is an important component because it often clouds our decision making and is something to take into account when empirically analyzing how humans and nonhumans interface with intelligent agents, and perhaps how they reciprocate actions without having emotionally grounded social behaviours.

3.6. Technical Investigations and Real-World Application

Of course, any intelligence-based technology would be useless if it could not interact with its users, including other intelligence-based technology, in an expected, precise, and ethical manner (Muehlhauser and Helm 2012; Laurent, Orseau and Armstrong 2016a; Soares, Yudkowsky, and Armstrong 2015b). Technical investigations are a marriage of conceptual and empirical investigations in that that they take what was learned and combine them to impart on a physical system all the important information gathered. The tripartite approach, however, doesn't limit technical investigations to occur in isolation. Often, changing the form factor of technology, especially considering intelligent agents can be housed in a variety of embodiments, may influence further conceptual or empirical investigations.

Further, there may be occasions where those additional investigations affect changes that are imperative to make to the technical aspects of the intelligent agent. Take the production of an intelligent agent in the form of a care robot that is to assist an elderly person with daily tasks—if at first designers believe that an assistant must ask users for permission to assist them with every micro task, they may soon learn that such requests impede the actual aid of the system and end up doing more to frustrate users than help them (van Wynsberghe 2013). Situations like this would call for further empirical investigations to tabulate the behaviour of both the care robots and the elderly people using it. Also, the situation may call for further conceptual investigations to gather more information about what the resultant technology is to do and should do.

Once again, we arrive at the fact that VSD can be of great service to producing intelligent agents—discussing technical investigations—given that intelligent agents include a broad scope of technologies to be used in various fields. Having the flexibility on offer means that changes can be made, even far into development cycles to ensure that all

three areas of the methodology are applied appropriately and consistently. The other benefit is that all three investigations, as detailed with the care robot example, feed into one another in fluid, well-connected manners; they are not separated by distinct design phases.

3.7. Avoiding After-The-Fact Dilemmas

Because intelligent agents can be used in diverse fields for a wide variety of cutting-edge technologies, there is a lot that can go wrong. Something unexpected can break, a feature can become outdated without a modulatory structure in place to modify it without a complete overhaul, something can function unintentionally, and worst of all, something terribly unsafe can occur. With all of these worries, there needs to be something to mitigate them, and that's where technical investigations help to prevent such mishaps.

As applies to all of VSD, it is not a perfect approach that can solve any engineering issue or account for every possible future consequence, but it can surely help to minimize the effects of several issues before they arise; it's much better than using an isolated ethical framework and resorting to ex-post facto remediation. Technical investigations provide the opportunity for the engineers and ethicists to experience the outcome of the conceptual and empirical studies—they can embody everything learned into an intelligent agent and then conduct further experiments and tests to see how the actual creation behaves within safe, controlled environments that replicate real-world scenarios.

And if something were to go awry, it can be modified, adjusted, accounted for, and accommodated in future iterations. The value of VSD really comes together with technical investigations because it pulls together important quantitative and qualitative data—it is the outcome of formal investigations, not a haphazard ad hoc attempt at alignment.

4. Harmonizing VSD

VSD begins from the central premise that technology is not value-neutral, meaning that technology is laden with values that are of ethical importance to individuals and society (Flanagan, C. Howe, and Nissenbaum 2008). Such moral values can include freedom, equality, trust, autonomy, or privacy, and each of these values affects and is affected by the technologies that embody them (Friedman 1997; Friedman and Kahn Jr. 2002). Other design methodologies tend to focus on the functionality and usability of innovations, whereas the VSD approach emphasizes the values that stakeholders (i.e., users, designers, corporate entities, etc.) hold to be important. In particular, VSD provides a grounded methodology that designers can levy when confronted with various and conflicting values (see Friedman and Kahn 2002). Because value-related issues are connected to the application of technology within a social context, VSD aims to incorporate those value solutions into the design and address any issues that may emerge during the early design phases before ubiquitous rollout.

Although numerous other design methodologies can be employed, as mentioned in the introduction to this chapter, the VSD approach does not seek to provide designers with an all-encompassing and exhaustive design methodology (Cummings 2006). Instead, VSD should be contextualized as a method that is meant to be harmonized with current practices in whichever field it is sought to be employed. Since the VSD

approach is similar to existent engineering approaches used by designers, the adoption of the VSD methodology in current research and development (R&D) practices makes it an attractive means by which to operationalize values in design (Cummings 2006). The harmonization of the VSD methodology to current practices requires a thorough knowledge of practices in the R&D of intelligent agents. By garnering a knowledge of current practice gaps, the shortcoming of existent methods can be uncovered after which VSD can employ its conceptual and empirical investigations to meet the needs of current practices.

The values of safety, efficacy, and proportionality serve as good point of departure. These values express the constructive benefits that responsibly designed intelligent agents can provide to society as well as begin to address many of the safety and social issues raised in scholarship regarding the development of unmitigated intelligent agents (Armstrong, Bostrom, and Shulman 2016; Baum 2016; Dodig Crnkovic and Çürüklü 2012).

An example of the issues emerging from the development of intelligent agents that patterns a clear instance of the design-for-values approach is the design and development of care robots. Throughout the design process value trade-offs must be weighed against the value of capability, efficiency, and safety (van Wynsberghe 2013). Capability would entail that the care robot is capable of providing a high standard of care by possessing qualities like strength, articulation, and intelligence in order to cater care to the need of the individual patient. Safety, on the other hand, would mean not causing harm to the patient either physically or through malfeasance. In trying to determine the moral weight solicited during these trade-offs, a means by which these values can be morally grounded is required. The VSD approach provides a transparent methodology through which designers can investigate the values of stakeholders, conceptualize any existent or emerging issues by drawing from the philosophical literature, and operationalize those values during the design phases.

Finally, the adoption of the VSD methodology can be enhanced by aligning existing testing practices with those proposed by the empirical investigation methods. Given the lack of universal regulations on intelligent agents, the tolerances and testing that they have to meet in order to be rolled out, as well as the lack of any agreed upon value-design landscape, a moral imperative to design intelligent agents with values becomes of the utmost importance. Intelligent agents as both an emerging and converging technology will almost certainly entail the emergence of new ethical and societal issues, as well as the exacerbation of current issues associated with its development. Integrating VSD with current practices could prove beneficial if the resulting amalgamation broadens existent practices. Firstly, the values that emerge as a result of the convergence characteristic of intelligent agents with other technologies now have a methodological framework in which they can be sufficiently evaluated for design. Secondly, the strict methodology would serve to address better critical issues associated not only with the technology itself, but with the activities of designers and developers, for instance by introducing potential surveillance techniques of AI designers in order to reduce the likelihood of rogue design (see Baum 2016).

Sufficiently applying the VSD methodology to the development of intelligent agents requires complex and highly specific insights on current design, developmental, and testing practices from all over the globe. Hence, experts from the field of AI and robotics research and development must be levied as integral parts of VSD's implementation for it to be tailored to the specific area of intelligent agent R&D. The last two decades of literature on the various potential applications of VSD serve as an excellent starting point for researchers and designers to draw from.

5. Conclusion

The introduction of intelligent systems and agents into the public domain poses new challenges for all societal sectors. New technological synergies that are powered by intelligent agents will experience the exacerbation of existent ethical issues while additionally giving rise to new ones. Likewise, this changing ethical landscape will affect the ways in which we tackle these new problems. In this chapter, we explored how Value-Sensitive Design (VSD) can provide a way in which designers can close the chasm between technological design and ethics both before and during the development of intelligent agents before they become ubiquitous.

The introduction of intelligent agents, embodied as autonomous physical entities or otherwise, will undoubtedly promote a level of ethical and social complexity that is common with other emerging and converging technologies such as nanotechnology (Drexler 1981, 2006; Phoenix and Drexler 2004), biotechnology (Diallo et al. 2012; MacGregor 2013; Tait and Levidow 1992) and ICT (Timmermans, Zhao, and van den Hoven 2011). Likewise, by framing intelligent agents as such, the inherent uncertainty and complexity can be employed as a sufficient starting point for design. The VSD approach provides a pragmatic framework suited towards the addressing of both existent and emerging ethical and social issues during all phases of technological design. VSD encourages and mandates the ethical evaluation of values, how those values are embodied in design, as well as how those integrated values actually play out in the application of developed innovations.

Although VSD has yet to be applied or precisely modulated for use in the design of intelligent agents, scholarship has begun to see the value in VSD and its potential applications to future technologies such as care robots (applies to Wynsberghe 2013), ICT (Friedman, Kahn Jr., and Borning 2006), and nanotechnology (Timmermans, Zhao, and van den Hoven 2011). Regardless, this chapter has provided a rough exploration of how VSD could be employed, and through such endeavours, has resulted in some preliminary results. First, it shows how most issues that are critical to design and important to stakeholders can be reduced to operational values as well as how stakeholders can come to be involved in the responsible design of innovations. Second, VSD can and should be integrated and take into account the existent design approaches employed by engineers and designers of intelligent agents. By bearing in mind and combining homogenously with current engineering and design practices, the adoption of the VSD methodology and all of its resulting boons will not only be easier but a more attractive option when it comes to designing with values.

The analyses and rudimentary approach of VSD outlined in this chapter are far from conclusive or exhaustive; further research and testing need to be undertaken to determine the long-term suitability of the VSD methodology to the design of intelligent agents. From an ethical standpoint, further conceptual research that investigates the issues and values implicated by intelligent agents is necessary, but more importantly, regarding the empirical and technical investigations, further research must be engaged with on how VSD can be harmonized into the concurrent and developing intelligent agent R&D practices.

6. References

Abney, Keith. 2014. “Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed.” In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by

Patrick Lin, Keith Abney, and George A. Bekey, 35–52. MIT Press.

Allen, Colin. 1997. "Calculated Morality : Ethical Computing in the Limit Colin." In *Cognitive, Emotive and Ethical Aspects of Decision Making and Human Action*, edited by Iva Smit and G Lasker, 1–5. Windsor: IIAS.

Armstrong, Stuart, Nick Bostrom, and Carl Shulman. 2016. "Racing to the Precipice: A Model of Artificial Intelligence Development." *AI and Society* 31 (2): 201–6. doi:10.1007/s00146-015-0590-y.

Baum, Seth D. 2016. "On the Promotion of Safe and Socially Beneficial Artificial Intelligence." *AI and Society*, no. July: 1–9. doi:10.1007/s00146-016-0677-0.

Bonnemains, Vincent, Claire Saurel, and Catherine Tessier. 2016. "How Ethical Frameworks Answer to Ethical Dilemmas: Towards a Formal Model." In *CEUR Workshop Proceedings*, 44–51.

Bourke, John, and Brian Duffy. 2003. "Emotion Machines: Projective Intelligence and Emotion in Robotics." *Cybernetic Intelligence, Challenges and Advances*, no. September: 20. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.112.2253
TeXampSymbrep=rep1
TeXampSymb#page=19.

Breazeal, Cynthia. 2003. "Emotion and Sociable Humanoid Robots." *International Journal of Human Computer Studies* 59 (1–2): 119–55. doi:10.1016/S1071-5819(03)00018-1.

Briggs, Pam, and Lisa Thomas. 2015. "An Inclusive, Value Sensitive Design Perspective on Future Identity Technologies." *ACM Transactions on Computer-Human Interaction* 22 (5): 1–28. doi:10.1145/2778972.

Clarkson, John, Simeon Keates, Roger Coleman, and Cherie Lebbon. 2003. *Inclusive Design: Design for the Whole Population*. Edited by John Clarkson, Simeon Keates, Roger Coleman, and Cherie Lebbon. London: Springer London. doi:10.1007/978-1- 4471-0001-0.

Cummings, Mary L. 2006. "Integrating Ethics in Design through the Value-Sensitive Design Approach." *Science and Engineering Ethics* 12 (4): 701–15. doi:10.1007/s11948-006-0065-0.

Denning, Tamara, Tadayoshi Kohno, and Henry M Levy. 2013. "A Framework for Evaluating Security Risks Associated with Technologies Used at Home." *CoMMuNiCatioNs of tHe aCM* 56 (1). doi:10.1145/2398356.2398377.

Diallo, Mamadou, Bruce Tonn, Pedro Alvarez, Philippe Bardet, Ken Chong, David Feldman, Roop Mahajan, Norman Scott, Robert Urban, and Eli Yablonovitch. 2012. "Implications: Convergence of Knowledge and Technology for a Sustainable Society." In *Convergence of Knowledge, Technology and Society*, edited by Mihail C. Roco, William S. Bainbridge, Bruce Tonn, and George Whitesides, 311–56. Springer International Publishing. doi:10.1007/978-3-319-02204-8_9.

Dodig Crnkovic, Gordana, and Baran Çürüklü. 2012. "Robots: Ethical by Design." *Ethics and Information Technology* 14 (1): 61–71. doi:10.1007/s10676-011-9278-2.

Doorn, Neelke. 2012. "Responsibility Ascriptions in Technology Development and Engineering: Three Perspectives." *Science and Engineering Ethics* 18 (1): 69–90. doi:10.1007/s11948-009-9189-3.

Drexler, K. Eric. 1981. "Molecular Engineering: An Approach to the Development of General Capabilities for Molecular Manipulation." *Proceedings of the National Academy of Sciences* 78 (9): 5275–78. doi:10.1073/pnas.78.9.5275.

———. 2006. "Engines of Creation 2.0. The Coming Era of Nanotechnology." *Anchor Books- Doubleday*, 576. doi:EB I DREX.

Flanagan, Mary, Daniel C. Howe, and Helen Nissenbaum. 2008. "Em-

bodying Values in Technology: Theory and Practice." In *Information Technology and Moral Philosophy*, edited by Jeroen van den Hoven and John Weckert, 322–53. New York, NY: Cambridge University Press. http://www.cambridge.org/catalogue/catalogue.asp?isbn=9780521855495&ss=cop.

Friedman, Batya. 1996. "Value-Sensitive Design." *Interactions* 3 (6): 16–23. doi:10.1145/242485.242493.

———. 1997. *Human Values and the Design of Computer Technology*. Edited by Batya Friedman. CSLI Publications. https://web.stanford.edu/group/cslipublications/cslipublications/site/1575860805.shtml#.

Friedman, Batya, and Peter H. Kahn Jr. 2002. "Value Sensitive Design: Theory and Methods." *University of Washington Technical*, no. December: 1–8. doi:10.1016/j.neuropharm.2007.08.009.

Friedman, Batya, Peter H. Kahn Jr., and Alan Borning. 2006. "Value Sensitive Design and Information Systems (PREPRINT)." *Human-Computer Interaction and Management Information Systems: Foundations*, 1–27. doi:10.1145/242485.242493.

Fussell, Susan R., Sara Kiesler, Leslie D. Setlock, and Victoria Yew. 2008. "How People Anthropomorphize Robots." In *Proceedings of the 3rd International Conference on Human Robot Interaction - HRI '08*, 145. doi:10.1145/1349822.1349842.

Goertzel, Ben. 2016. "Infusing Advanced AGIs with Human-Like Value Systems : Two Theses." *Journal of Evolution and Technology* 26 (1): 50–72.

Hellström, Thomas. 2013. "On the Moral Responsibility of Military Robots." *Ethics and Information Technology* 15 (2): 99–107. doi:10.1007/s10676-012-9301-2.

Hoven, Jeroen van den, and Klaus Jacob. 2013. *Options for Strengthening Responsible Research and Innovation*. doi:10.2777/46253.

Hoven, Jeroen Van den, Gert Jan Lokhorst, and Ibo Van de Poel. 2012. "Engineering and the Problem of Moral Overload." *Science and Engineering Ethics* 18 (1): 143–55. doi:10.1007/s11948-011-9277-z.

Johnson, Aaron M., and Sidney Axinn. 2013. "The Morality of Autonomous Robots." *Journal of Military Ethics* 12 (2): 129–41. doi:10.1080/15027570.2013.818399.

Keates, Simeon, and John Clarkson. 2003. "Countering Design Exclusion." In *Inclusive Design: Design for the Whole Population*, edited by John Clarkson, Simeon Keates, Roger Coleman, and Cherie Lebbon, 438–53. London: Springer London. doi:10.1007/978-1-4471-0001-0_27.

Laurent Orseau, and Stuart Armstrong. 2016a. "Safely Interruptible Agents." In *32nd Conference on Uncertainty in Artificial Intelligence.*

———. 2016b. "Safely Interruptible Agents." In *32nd Conference on Uncertainty in Artificial Intelligence.* https://intelligence.org/files/Interruptibility.pdf.

MacGregor, Donald. 2013. "Convergence Platforms: Human-Scale Convergence and the Quality of Life." *Convergence of Knowledge,Technology and Society:Beyond Convergence of Nano-Bio-Info-Cognitive Technologies*, no. July: 1–52. doi:10.1007/978-3- 319-02204-8_2.

Mermet, Bruno, and Gaele Simon. 2016. "Formal Verification of Ethical Properties in Multiagent Systems." In *CEUR Workshop Proceedings*, 26–31.

Muehlhauser, Luke, and Nick Bostrom. 2014. "Why We Need Friendly Ai." *Think* 13 (36): 41–47. doi:10.1017/S1477175613000316.

Muehlhauser, Luke, and Louie Helm. 2012. "Intelligence Explosion and Machine Ethics." *Singularity Hypothesis: A Scientific and Philosopohical Assessment* 6: 1–28. doi:10.1007/978-3-642-32560-1_6.

Nadeau, Joseph E. 2006. "Only Androids Can Be Ethical." In *Thinking about Android Epistemology*, edited by K. M. Ford, C. N. Glymour, and P. J. Hayes. AAAI Press

(American Association for Artificial Intelligence).

Newell, Alan F., P. Gregor, M. Morgan, G. Pullin, and C. Macaulay. 2011. "User-Sensitive Inclusive Design." *Universal Access in the Information Society* 10 (3): 235–43. doi:10.1007/s10209-010-0203-y.

Phoenix, Chris, and K. Eric Drexler. 2004. "Safe Exponential Manufacturing." *Nanotechnology* 15 (8): 869–72. doi:10.1088/0957-4484/15/8/001.

Rolf, Matthias, and Nigel Crook. 2016. "What If: Robots Create Novel Goals? Ethics Based on Social Value Systems." In *CEUR Workshop Proceedings*, 20–25.

Scheutz, Matthias, and Bertram F. Malle. 2014. " 'Think and Do the Right Thing' - A Plea for Morally Competent Autonomous Robots." In *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, ETHICS 2014*. doi:10.1109/ETHICS.2014.6893457.

Soares, Nate. 2016. "The Value Learning Problem." In *Ethics for Artificial IntelligenceWorkshop at 25th International Joint Conference on Artificial Intelligence*, 1–8. Machine Intelligence Research Institute. http://intelligence.org/files/ValueLearningProblem.pdf.

Soares, Nate, and Benya Fallenstein. 2014a. "Agent Foundations for Aligning Machine Intelligence with Human Interests : A Technical Research Agenda." doi:10.1007/978-3-662-54033-6_5.

———. 2014b. "Agent Foundations for Aligning Machine Intelligence with Human Interests : A Technical Research Agenda," 1–14. doi:10.1007/978-3-662-54033-6_5.

Soares, Nate, and Benjamin A. Levinstein. 2017. "Cheating Death in Damascus." In *14th Annual Formal Epistemology Workshop*, 1–20. Machine Intelligence Research Institute. https://intelligence.org/files/DeathInDamascus.pdf%0Ahttps://intelligence.org/2017/03/18/new-paper-cheating-death-in-damascus/.

Soares, Nate, Eliezer Yudkowsky, and Stuart Armstrong. 2015a. "Corrigibility." In *AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Publications. http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10124.

———. 2015b. "Corrigibility." In *AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Publications.

Sotala, Kaj. 2016. "Defining Human Values for Value Learners." *AAAI-16 AI, Society and Ethics Workshop*, 113–23.

Stahl, Bernd Carsten. 2004. "Information, Ethics, and Computers: The Problem of Autonomous Moral Agents." *Minds and Machines* 14 (1): 67–83. doi:10.1023/B:MIND.0000005136.61217.93.

Stocker, Michael. 1976. "The Schizophrenia of Modern Ethical Theories." *Journal of Philosophy* 73 (14): 453–66. doi:10.2307/2025782.

Tait, Joyce, and Les Levidow. 1992. "Proactive and Reactive Approaches to Risk Regulation. The Case of Biotechnology." *Futures* 24 (3): 219–31. doi:10.1016/0016-3287(92)90032-B.

Tao, Zhang, Zhu Biwen, Lashanda Lee, and David Kaber. 2008. "Service Robot Anthropomorphism and Interface Design for Emotion in Human-Robot Interaction." In *4th IEEE Conference on Automation Science and Engineering, CASE 2008*, 674–79. doi:10.1109/COASE.2008.4626532.

Taylor, Jessica, Eliezer Yudkowsky, Patrick Lavictoire, and Andrew Critch. 2016a. "Alignment for Advanced Machine Learning Systems." *Machine Intelligence Research Institute*. https://intelligence.org/files/AlignmentMachineLearning.pdf.

———. 2016b. "Alignment for Advanced Machine Learning Systems." *Machine In-*

telligence Research Institute.

Timmermans, Job, Yinghuan Zhao, and Jeroen van den Hoven. 2011. "Ethics and Nanopharmacy: Value Sensitive Design of New Drugs." *NanoEthics* 5 (3): 269–83. doi:10.1007/s11569-011-0135-x.

Torrance, Steve. 2013. "Artificial Agents and the Expanding Ethical Circle." *AI and Society* 28 (4): 399–414. doi:10.1007/s00146-012-0422-2.

Waser, Mark R. 2016. "Implementing a Seed Safe/Moral Motivational System with the Independent Core Observer Model (ICOM)." *Procedia Computer Science* 88: 125–30. doi:10.1016/j.procs.2016.07.415.

Wiltshire, Travis J. 2015. "A Prospective Framework for the Design of Ideal Artificial Moral Agents: Insights from the Science of Heroism in Humans." *Minds and Machines* 25 (1): 57–71. doi:10.1007/s11023-015-9361-2.

Wolf, Susan. 1983. "Moral Saints." *The Journal of Philosophy* 80 (10): 563–70. doi:10.2307/2026571.

Wynsberghe, Aimee van. 2013. "Designing Robots for Care: Care Centered Value-Sensitive Design." *Science and Engineering Ethics* 19 (2): 407–33. doi:10.1007/s11948-011-9343-6.

Acknowledgements

We thank Tony Barrett and Phil Torres for feedback on an earlier draft. Any errors are the authors' alone. The views in the chapter 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.

Angelo Frank De Bellis Angelo De Bellis is technical communicator and independent researcher with interests in the ethics of artificial intelligence and the philosophy of technology