**Designing AI for Explainability and Verifiability: A Value Sensitive Design Approach to Avoid Artificial Stupidity in Autonomous Vehicles**

Steven Umbrelloa and Roman V. Yampolskiyb

a Institute for Ethics and Emerging Technology; University of Turin, Italy [steven.umbrello@unito.it](mailto:steven.umbrello@unito.it)

b University of Louisville, Louisville, roman.yampolskiy@louisville.edu

**Abstract**

One of the primary, if not most critical, difficulties in the design and implementation of autonomous systems is the black-boxed nature of the decision-making structures and logical pathways of autonomous systems. For this reason, the values of stakeholders become of particular significance given the risks posed by opaque structures of intelligent agents (IAs). This paper proposes the Value Sensitive Design (VSD) approach as a principled framework for incorporating these values in design. The example of autonomous vehicles is used as a case study for how VSD offers a systematic way for engineering teams to formally incorporate existing technical solutions towards ethical design, while simultaneously remaining pliable to emerging issues and needs. It is concluded that the VSD methodology offers at least a strong enough foundation from which designers can begin to anticipate design needs and formulate salient design flows that can be adapted to changing ethical landscapes.

**Keywords**: value sensitive design, artificial intelligence, autonomous vehicles, explainability, verifiability, applied ethics

1. **Introduction**

One of the primary, if not most critical, difficulties in the design and implementation of autonomous systems is the black-boxed nature of the decision-making structures and logical pathways of autonomous systems. For this reason, the values of stakeholders become of particular significance given the risks posed by opaque structures of intelligent agents (IAs). This paper proposes the Value Sensitive Design (VSD) approach as a principled framework for incorporating these values in design and retaining meaningful human control over them (Santoni de Sio & van den Hoven, 2018).

VSD is often described as a principled approach to technological design, one that aims to incorporate and account for the values of various stakeholder groups early on and throughout the design process (Friedman, 1997). It begins with the premise that technology is not value-neutral, but instead is sensitive to the values held by stakeholders, such as the designers, engineers, and users among others (Friedman & Kahn Jr, 2003; Santoni de Sio & van den Hoven, 2018; van den Hoven & Manders-Huits, 2009).

To the best of our knowledge, this is the first paper to evaluate the applicability of the VSD approach to AI as it pertains to the values of explainability and verifiability (Yampolskiy, 2017). Prior literature on VSD has focused on its methodological foundations (Friedman, Hendry, & Borning, 2017; Friedman, Kahn, Borning, & Huldtgren, 2013; Umbrello, 2018), its applicability to existing technologies such as energy systems and care robotics (Oosterlaken, 2015; van Wynsberghe, 2012, 2016), as well as its applicability to AI in general and other advanced technologies such as molecular manufacturing (Timmermans, Zhao, & van den Hoven, 2011; Umbrello, 2019b, 2019a; Umbrello & De Bellis, 2018). These studies provide useful information on the VSD approach in general, as well as why and how it can and perhaps should be used for the development of IAs. However, none of these focus on the values of particular interest for the safe development of beneficial IAs. Similarly, unlike other research projects which focus on technologies as concepts, rather than particular examples, this project takes up autonomous vehicles (AVs) as a case study to demonstrate practical means that designers can adopt in designing IAs with the values of explainability and verifiability in mind (among others).

Section 2 outlines both some of the current difficulties and issues that arise from the development of AVs as well as how the above stated values come in to play in their development. Similarly, a brief discussion and justification are given for the choice of using VSD as a design approach rather than other design-for-values methodologies. Section 3 outlines the VSD methodology in full, giving particular emphasis to empirical and technical investigations. Section 4 provides a cursory account of how explainability and verifiability can be balanced in design requirements for IAs in general. Section 5 discusses how those design requirements can be better understood in the case of AVs. Section 6 concludes this paper by summarizing its findings as well as by providing suggestions for potentially fruitful future research.

1. **Emerging Issues with Autonomous Vehicles**

Discussions on autonomous systems both in military and civil spheres have held centre stage in applied ethics circles and scholarship. This is for good reason, the exponential advancements of technical systems such as neural networks, machine learning, robotics, and sensor technologies permit a fertile ground for autonomous systems to proliferate across domains with decreasing need for human command and control. Given the abdication of control to these systems, the primary issue that emerges as a result of their introduction into society surrounds questions of responsibility and liability. The placement of responsibility on human actors becomes contentious given that responsibility has traditionally implicated notions of autonomy, but if autonomy is held by nonhuman systems, where does the burden of responsibility lie? What if an AV kills a pedestrian going about their day on the sidewalk and the designers of that AV’s programming cannot discern the decision pathway of the logic that lead to the AV’s decision to swerve onto the sidewalk? Can we reasonably punish the AV given that it is functionally autonomous? And can we, with a clear conscious, put the blame on the designers, who them themselves never programmed such a set of inputs? These are some of the basic questions that have persisted in the literature on autonomous system, particularly within the legal fields that are interested with legislation and liability issues surrounding AV’s which are already present on many roads (Contissa, Lagioia, & Sartor, 2017b, 2017a; Dogan et al., 2016; Thornton, Lewis, Zhang, Kochenderfer, & Gerdes, 2018).

Roads are continuously occupied by a variety of stakeholder groups such as the vehicle drivers themselves, pedestrians, and cyclists. The implied values that are held by these stakeholders in large part govern their actions, reactions and overall expectations to situations that can, and often do, arise on roads. The introduction and continual transition towards autonomation, in that case, on the roads with AVs, will most likely also be predicated by similar roadway values. Some of these existent roadway values are safety and permissibility of road rules. Naturally, emerging values such as security and privacy, which can easily come into conflict with one another, emerge with AVs. System designers are burdened with the task of determining how these human values can be translated into design requirements that can be embedded into an autonomous system. One of the basic ways that designers can do this is by consulting stakeholders and integrating elicited values into the design of the decision matrix algorithms (DMAs) of AVs.

DMAs are one of the primary algorithms used to form the machine learning set of AVs (Gupta, 2017; Leben, 2017; Wachter, Mittelstadt, & Floridi, 2016). A popular example of a DMA is adaptive boosting (AdaBoosting), other employed algorithms include clustering algorithms such as K-means, pattern recognizers (classifiers), support vector machines, and regression algorithms such as neural network regression. DMAs are chosen over the other implicated, and no less important algorithm systems because DMAs functions by methodically evaluating and ranking the efficacy of the relation between data sets and values. Because of this function, DMAs are employed for primary decision making. Every action the car takes, whether it be to accelerate, brake suddenly, or swerve is predicated on the strength of the rank-relationship attributed to the recognition and movement of environmental entities based on sensor input and predictive analysis of that data. The rank-ordering of these relationships is a function of the independent training of models that are then aggregated to create a predictive decision-making system with the goal to reduce errors in judgement. The training of these models with chosen data sets itself implicate values (i.e., which are chosen, opportunity costs of chosen inputs), provides the perfect place for designers to directly intervene with intention of actively guiding the formation of these models through principled stakeholder engagement.

The VSD approach is chosen because it is founded on the premise that technology is something that is value-laden and thus is of significant ethical importance (Davis & Nathan, 2014; Friedman et al., 2015; Pitt & Diaconescu, 2016). VSD is a principled approach to design that is divided into three distinct investigations (Figure 1), referred to commonly as a ‘tripartite methodology’, conceptual investigations, empirical investigations, and technical investigations (Friedman & Kahn Jr., 2003; Friedman et al., 2013). The framework is designed to be iterative and recursively self-reflective as it aims to continually weigh the values of both direct and indirect stakeholder groups throughout the design process. Because of this, the VSD methodology aims to be, and is typically employed in technological design where human values come into conflict with each other and where the design of such technologies implicate considerable ethical concern (van den Hoven, Lokhorst, & van de Poel, 2012). Similar to how Friedman, Kahn, and Borning (Friedman et al., 2013) used VSD in the realm of human-computer interaction to show how the values of privacy and usability needed to be balanced, Umbrello (2019b) showed that stakeholders implicated in the development of AI in the United Kingdom prized the values of transparency, control, data privacy, and security, values in which need to be delicately balanced through stakeholder coordination and cooperation.

**Figure 1.** The recursive VSD tripartite framework. Source: (Umbrello, 2020b).

Regardless, the applicability of the VSD methodology to a variety of technological artefacts makes its acceptance by design groups attractive because various scholars have demonstrated its ability to be adapted to the artefact at hand and to be streamlined into existing design practices. For example Timmermans et al. (Timmermans et al., 2011) adopt the VSD approach for the design of nanopharmaceuticals by adopting the values existent in the medical field and van Wynsberghe (van Wynsberghe, 2013) similarly draws from the values of care to modify the VSD methodology for application to care robots.

In the design of the models that form DMAs, designers already program with the values of safety and efficacy in mind. They are the two primary values that are most commonly sought via current design, for example, Chen, Peng, and Grizzle (2018) presents a design for an obstacle avoidance algorithm for low-speed autonomous vehicles with the intent of balancing efficacy by reducing control effort while of course prioritizing safety by minimising pedestrian and obstacle collision. Similarly, Kamali et al. (2017) propose formal verification (FV) of AV code viz. program model-checking algorithms to increase the safety of AV platooning (i.e., organization of several AVs into convoys or platoons). These examples demonstrate the current attempts by designers to integrate important human values into the design of AVs. Yet, the consistent attempt at balancing the values of safety and efficacy leads to what VSD theorists call *moral overload* since both of the values are prized and implicated by enrolled stakeholders, yet they are often at odds with each other when formulated as technical design requirements (van de Poel, 2017; van den Hoven et al., 2012). Similarly, aside from these two values, other values such as trust and control are implicated in AV design which can similarly come into conflict (Abraham et al., 2017; Contissa et al., 2017b; Prokhorov, 2018; Wang, Gong, Zhou, Li, & Peeta, 2018; Yan, Xu, & Liu, 2016). For this reason, it would be useful for designers to adopt a principled approach to design that provides the tools to account for moral overload as well as to adjudicate and balance *prima facie* conflicting moral values in a way that best satisfies what stakeholders value.

This paper proposes that VSD can help bridge the chasm in the design process of DMAs for AVs. Not only this, but a more comprehensive set of values is considered that may be implicated in AV design aside from safety and effectiveness – particularly explainability (i.e., transparency) and verifiability - as well as proposals for how moral overload can be resolved through certain design flows. This paper thus aims to demonstrate an adapted form of the VSD approach for DMA model training with particular emphasis on obstacle avoidance models. In order for AVs to be deployed, their decisions – a function of model training for DMA attribution – require FV to ensure transparent lines of decision making logic that can be explained to/by designers to minimize incidents on the road. This FV policy is designed using VSD as a framework. This *value-sensitively designed* DMA, although limited to the AVs controller-code and not to the full model of the autonomous system, nonetheless captures the decision-making structure of the agents code that ensures it does not violate designed values such as safety.

The following section outlines in greater depth the VSD methodology, highlighting its tripartite structure of conceptual investigations, empirical analysis, and technical limits and constraints.

1. **Value Sensitive Design**

The VSD methodology is traditionally conceived of as a tripartite methodology consisting of three stages of investigations: conceptual, empirical and technical (Friedman et al., 2017). The first of the three, conceptual investigations, consists of answering the following questions: Who are the stakeholders? What are the related values related to the technology in question? Where do certain parameters begin and end when discussing the bounds of usability versus conflicting values like transparency and privacy or safety and efficacy? Who are the 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 theoretical and philosophical questions fall within the purview of conceptual investigation (Denning, Kohno, & Levy, 2013).

The second phase, empirical investigations, aims to use both quantitative and qualitative tools and analyses to determine if the distilled conceptual values can meet the need of stakeholders in design. Things like statistical data that describes patterns of human behavior, assessments that measure 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 (Umbrello & De Bellis, 2018). This stage ultimately aims to determine if the design of a technology maps onto the conceptual results, if not, a recursive feedback back to conceptual investigations is needed to determine how those values can be better mapped onto design.

Finally, technical investigations looks at the technical limitations of the artefact in question. Because certain technologies and materials can support or constrain certain values, these investigations aim to determine how the actual technical specifications of a design can be best tailored to support the values of stakeholders while limiting unwanted or potentially emergent problems. The technical questions become important in the operationalization of values given that they can constrain how those values are instantiated in the design (Friedman & Kahn Jr., 2002).

In total, the three parts are meant to be iterative, feeding into one another until alignment between the three can be harmonized. Designers tend to already engage in self-feedback and redesign until they meet their desired criteria, the VSD methodology enables a more principled way to go about formalizing this otherwise implicit practice to better ensure value-alignment both in the early phases and throughout the design process (Friedman & Hendry, 2019; Umbrello, 2020a).

1. **Applying VSD and the *Belief-Desire-Intention* (BDI) Model**

In beginning the VSD process, one of the most critical steps involved is determining both the direct and indirect stakeholders involved. In the case of AVs, it requires tracing the development pathway from origin to use. Direct stakeholders can be the designers and engineers of AVs themselves, whether they be mechanical engineers or computer scientists responsible for system programming. Users i.e., the drivers (and occupants) of the AVs themselves are naturally enrolled as direct stakeholders. Similarly, the industries responsible for commissioning such vehicles and the public at large, particularly pedestrians can be considered indirect stakeholders. Although the roles between direct and indirect stakeholdership is dynamic, and is contingent on the scenarios under consideration. Tracing this development-use pathway is useful for seeing who is enrolled in the design process and how they can be further implicated in determining the values that are important to them. Various methodologies within VSD are apt to stakeholder discovery and elicitation such as stakeholder analysis (Czeskis et al., 2010), stakeholder tokens (Yoo, 2017) and Envisioning Cards (Friedman & Hendry, 2012).

When taking these stakeholders under consideration in any particular scenario, it becomes crucial to distill the relevant values at play. These values are obviously significant because they are always-already implicated in design (Pinch & Bijker, 1987; Winner, 2003). What becomes important is to highlight which values are implicated and how desired values can be supported among other values that may be in tension (i.e., privacy and security) (van den Hoven et al., 2012). Tools such as value source analysis (Borning, Friedman, Davis, & Lin, 2005), value-oriented coding manual (Kahn Jr, Friedman, Freier, & Severson, 2003) and/or value-orient mock-up or protypes (Woelfer & Hendry, 2009) can be used towards this end. As already mentioned, traffic schemes typically include the human values of safety and lawfulness. Through a conceptual investigation of stakeholders involved further values can be distilled such as trust, autonomy, transparency and privacy (i.e., Umbrello, 2019b). Because each of these values can lead to different design requirements and flows, it becomes important how they are conceptualized, balanced and translated as engineering goals. Van de Poel (2013) uses a value hierarchy to help designers translate value through norms and into tangible design requirements (Figure 2). This paper focuses on the value of transparency given that it implicates and often comes into conflict with other important values in AVs and AI systems in general such as privacy, safety and efficiency.

**Figure 2.** Values hierarchy. Source: (Van de Poel, 2013).

Transparency is often cited as a good in itself in much of the AI literature (Johri & Nair, 2011; Mortier, Henderson, McAuely, & Crowcroft, 2014; Vermaas, Hekkert, Manders-Huits, & Tromp, 2014). However, the term is sensitive to context, and is far more nuanced than is often considered. Although often beneficial within the context of algorithmic verifiability and understanding, there are cases where transparency may not be so good for stakeholders. The efficacy of transparency is similarly context dependent on goals and definition. It can take the form of designers being able to determine how well a system is working or not and how it can be improved where it needs to be, helping publics to understand the strengths and drawbacks of a particular system to encourage trust and similarly to encourage users and designers to be able to anticipate future actions of a system, to trace a decision stream of a system in the event of an error (or an accident in the case of AVs) and to attribute cause and responsibility (Mecacci & de Sio, 2019). This is not an exhaustive list of way to conceptualize transparency. What those listed have in common is that transparency is construed as a general benefit to the society at large. Of course what determines the strength of, and makes them generally beneficial hinges on an ambiguous account of authenticity in the information provided to users and programmers and that nothing is absconded that may be crucial to agents implicated.

Still transparency can similarly come into tension with other important values, particularly when considering AVs. The issues arises from the meta-consideration of construing transparency as a design goal, rather than a means to support or limit other design requirements. For example, full transparency as a mandated requirement, such as mandating that the source code of systems be fully open, can lead to not only the manipulation of such code, but can also disincentivize industry leaders to innovate given the lack of proprietorship (Ghani, 2016; Roco, 2008).

However, perhaps the most obvious tension that the value of transparency can have is in contrast to the value of privacy. Many individuals consider a basic right to privacy as fundamental, and thus should limit the amount of transparency that is implemented in a system. Designers can come into tension here where stakeholders want both a right to data privacy but also transparency of how a system functions. The tension is most obvious where greater transparency can lead to greater trust in a system (where the interests between the system and the stakeholder align) but also where privacy in a system also fosters a stakeholders trust to use it. Both the values of transparency and safety, although in tension with one another, can foster other values (i.e., trust, confidence) but in different ways. Because of this tension, and because of transparency’s importance in the deployments of AVs we should take care not to conflate transparency as a goal per se, but rather as a means of supporting or constraining other important values, as a design flow. In a word, transparency as an instrumental value, a value-in-process is how it should be conceptualized rather than as an end-value (Boscoe, 2019).

One of the initial ways of conceptualizing these design flows through VSD and for engineers to accept and adopt the methodology is by integrating the methodology with similar practices and theories, in this case, with those particular to AVs. For example, a *rational agent* paradigm can be adopted as a hybrid architecture for verification needs given that it permits discrete and continuous control systems to be separated, yet verified in greater depth (Kamali et al., 2017; Wooldridge, 2002). This level of transparency promotes safety given that each level of discrete decision making is discernable to the engineers and allows them to guide decision making towards exclusively safe ends. So not only can programmers see *what* an AV chooses to do in a given scenario, but also *why* it chooses to do so (Fisher, Dennis, & Webster, 2013). Not only does this rational agent approach dissolve opacity, but it promotes self-improving design flows which in turn promotes its acceptance by engineering teams and industry.

Hence, the models from which DMA attribution models can begin to be conceptualized can start from rational agent paradigms. One of the most generally adopted models for both conceptualizing these types of rational agents as well as executing them in the engineering space is through the *Belief-Desire-Intention* (BDI) model (Caillou, Gaudou, Grignard, Truong, & Taillandier, 2017; Cointe, Bonnet, & Boissier, 2016; Lee & Son, 2008; Rao & Georgeff, 1992). A BDI modeled agent is characterized explicitly by its appellation: its beliefs, desires, and intentions. *Beliefs* are the agent’s impression of the external world, its *desires* are its end-goals that are to be captured, and its *intentions* are the agent’s concurrent actions in-progress towards its desires. DMA agents modeled with the BDI framework have a finite set of scenario parameters, how an agent behaves is constrained by its beliefs and associated end-goals. Similarly, an event succession of both sensor inputs and resultant beliefs are stored. Naturally, a model such as this provides various advantages for AVs as well as autonomous systems in general. The first is that it structurally separates response controllers from the high-level decision-making systems, this promotes that ability to discern the high-level reasoning structure and formally verify the decisions taken as well as strongly demarcates scenario selection and scenario execution (Dennis, Fisher, Lincoln, Lisitsa, & Veres, 2016). This transparent hierarchical structure promotes value-laden scenario programming of the model, supporting certain choice structures based on conceptual requirements distilled during initial VSD investigations.[[1]](#footnote-1)

Beginning with this transparently hierarchical structure, DMAs can be implemented in AV systems as an initial means by which designers can begin to conceptualize a value-sensitive approach to AV design. Research into this exact area has already been undertaken in platooning research where a ‘platoon’ or convoy of AVs synchronically follow a lead vehicle that is under human control (Hendrickson & Van Nieuwstadt, 2018; Kamali, Linker, & Fisher, 2018). The proposal to have platooned AV’s designed with DMA’s based on BDI models preserves meaningful human control of these autonomous systems, despite the lack of full autonomy and the adaptive behavior to novel sensor stimuli that characters machine learning algorithm approaches (Calvert, Mecacci, Heikoop, & de Sio, 2018).

The following discussion section outlines some implications of this approach as well as its limitations and potential further research avenues that may prove fruitful for full autonomous vehicles.

1. **Discussion**

There are numerous drawbacks and as such, many fruitful areas of potential future research that this paper implicates. Firstly, the rigid structure of the BDI hierarchy and DMA control system in general does naturally preclude any built-in learning, planning, and adaptation from environmental inputs or past events. What this does then is preclude, similarly, machine learning systems that have become desirable given their ability to adapt, learn from past experience and make decisions given novel input. What the proposal of this paper implicates then is a an explicit modeling structures that promotes transparency as a means towards enhancing safety and operability of AVs. Similarly this can build trust with the designers and users alike in their ability to understand the rational decisions taken by the agent given a certain set of input parameters in the models. Further research then should look at how the VSD methodology can balance the value requirements distilled by design teams while considering advanced machine learning systems. Whether this is even feasible is not the subject of this paper, but may prove to be rewarding as the harmonization between machine learning systems and their ability to adapt to changing inputs while remaining aligned with stakeholder values seems to predicate obvious boons.

Secondly, the value tensions that arise with different conceptualizations of what transparency is and how it is construed in engineering practice should be more closely considered. One way to conceptualize transparency other than the traditional *per se* virtue of it, is to look at it as a design flow that can be used to guide engineers and programmers to conceptualizing other important values where transparency can be used as a way to either support to curtail those values in design (i.e., as an instrumental value). Not only this, but transparency can be expounded in another way. The value of transparency is typically construed in the *human🡪system* direction where it is understood as the ability for the human (designer, engineer, programmer, users, etc.) to understand the what, why and how of a system’s decisions. However, future research should look at the transparency dynamics of the *system🡪human* relation where human actions become transparent to the system. A perfect example in the case of AVs is when a human pedestrian waves the car to proceed. It will become particularly constructive to consider the transparency of human actions and motivations in this respect (Calvert et al., 2018; Santoni de Sio & van den Hoven, 2018). This can be extended similarly to other AI systems such as autonomous weapons systems and the ability for those systems to understand non-verbal commands given by friendly combatants, or even non-friendly ones such as in cases where enemy combatants or civilians surrender (Johnson & Axinn, 2013; Klincewicz, 2015; Umbrello, Torres, & De Bellis, 2020).

Similarly, another fruitful avenue for further research would be in the transparency and interpretability dynamics of machine-machine relationships. Steps forward in this direction have been taken to look at how autonomous systems can communicate, coordinate and execute tasks together (Mermet & Simon, 2016; Mordatch & Abbeel, 2018; Tampuu et al., 2017). The organization and dynamics of these multi-agent ensembles should be further explored for a number of reasons. The first would be the cooperation between differing systems to autonomously institute extensible concepts that can apply generally goes beyond the simply linear transmission of narrow information. The benefit of this is the efficient communication of hierarchical concepts that are adaptable and thus can be utilized more generally. Naturally, it will remain important for human to retain meaningful control over these autonomous hierarchies, but it may be even more critical, and simpler, for designers to focus on machine-machine communications as a starting point. Where VSD researchers should look is ways to retain a level of interpretability of machine-machine cooperation structures so that complexity of hierarchy and communication developments do not become opaque over time.

Finally, an important area that is predicated by explainability and verifiably is on the very concept of human interpretability. How is interpretability measured and under what parameters is it satisfied when considering autonomous systems? Perhaps one pragmatic way to move forward on this would be to simply consider performance attributes rather than trying to empirically quantify internal explainability per se which comes with a host of issues attached (Bau, Zhou, Khosla, Oliva, & Torralba, 2017; Kim et al., 2018). Further research in this area for both external performance metrics as well as a more holistic understanding of internal comprehension may prove beneficial to long-term value-based AI development.

1. **Conclusions**

In this humble attempt at a VSD application towards AVs, we demonstrate one of the possible ways to formalize the approach into existing engineering practices. Conceptual and technical investigations of the VSD were highlighted as the most explicit areas where designers can draw from to formally connect human values to design requirements. This paper proposed that the decision matrix algorithms of AVs provide a potentially fruitful starting point for how values can be implemented in design through the training and programming of models. Because of this, engineers are conceived as designers that work throughout the design process of AVs and work directly with stakeholder groups. Because of this, the VSD methodology directly enrolls not only publics and industry, but also policy leaders and legislators as stakeholders that can co-create technologies. Further research could take the form of how to formally engage with policy leaders as stakeholders during the early phases and throughout the design process so that policy and technology can harmoniously co-constitute one another.

**Acknowledgements**

Any remaining errors are the author’s alone. The views in the paper are the authors’ alone and not the views of the Institute for Ethics and Emerging Technologies.

**References**

Abraham, H., Lee, C., Brady, S., Fitzgerald, C., Mehler, B., Reimer, B., & Coughlin, J. F. (2017). Autonomous vehicles, trust, and driving alternatives: A survey of consumer preferences. In *Transportation Research Board 96th Annual Meeting, Washington, DC* (pp. 8–12).

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. *ArXiv Preprint ArXiv:1704.05796*.

Borning, A., Friedman, B., Davis, J., & Lin, P. (2005). Informing public deliberation: Value sensitive design of indicators for a large-scale urban simulation. In *ECSCW 2005* (pp. 449–468). Springer.

Boscoe, B. (2019). Creating Transparency in Algorithmic Processes. *Delphi - Interdisciplinary Review of Emerging Technologies*, *2*(1). Retrieved from https://doi.org/10.21552/delphi/2019/1/5

Caillou, P., Gaudou, B., Grignard, A., Truong, C. Q., & Taillandier, P. (2017). A Simple-to-use BDI architecture for Agent-based Modeling and Simulation. In *Advances in Social Simulation 2015* (pp. 15–28). Springer.

Calvert, S. C., Mecacci, G., Heikoop, D. D., & de Sio, F. S. (2018). Full platoon control in Truck Platooning: A Meaningful Human Control perspective. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (pp. 3320–3326). IEEE.

Chen, Y., Peng, H., & Grizzle, J. (2018). Obstacle Avoidance for Low-Speed Autonomous Vehicles With Barrier Function. *IEEE Transactions on Control Systems Technology*, *26*(1), 194–206. https://doi.org/10.1109/TCST.2017.2654063

Cointe, N., Bonnet, G., & Boissier, O. (2016). Multi-agent based ethical asset management. In *CEUR Workshop Proceedings* (pp. 52–57).

Contissa, G., Lagioia, F., & Sartor, G. (2017a). Accidents Involving Autonomous Vehicles: Legal Issues and Ethical Dilemmas. *JUSLETTER*, 1–7.

Contissa, G., Lagioia, F., & Sartor, G. (2017b). The Ethical Knob: ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, *25*(3), 365–378. https://doi.org/10.1007/s10506-017-9211-z

Czeskis, A., Dermendjieva, I., Yapit, H., Borning, A., Friedman, B., Gill, B., & Kohno, T. (2010). Parenting from the pocket: Value tensions and technical directions for secure and private parent-teen mobile safety. In *Proceedings of the Sixth Symposium on Usable Privacy and Security* (p. 15). ACM.

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

Denning, T., Kohno, T., & Levy, H. M. (2013). A framework for evaluating security risks associated with technologies used at home. *Communications of the ACM*, *56*(1). https://doi.org/10.1145/2398356.2398377

Dennis, L. A., Fisher, M., Lincoln, N. K., Lisitsa, A., & Veres, S. M. (2016). Practical verification of decision-making in agent-based autonomous systems. *Automated Software Engineering*, *23*(3), 305–359.

Dogan, E., Chatila, R., Chauvier, S., Evans, K., Hadjixenophontos, P., & Perrin, J. (2016). Ethics in the design of automated vehicles: The AVEthics project. In *CEUR Workshop Proceedings* (pp. 10–13).

Fisher, M., Dennis, L., & Webster, M. (2013). Verifying autonomous systems. *Communications of the ACM*, *56*(9), 84–93.

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., & Hendry, D. G. (2012). The envisioning cards: A toolkit for catalyzing humanistic and technical imaginations. In *Proceedings of the 30th International Conference on Human Factors in Computing Systems - CHI ’12* (pp. 1145–1148). https://doi.org/10.1145/2207676.2208562

Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. Cambridge, MA: Mit Press.

Friedman, B., Hendry, D. G., & Borning, A. (2017). A Survey of Value Sensitive Design Methods. *Foundations and Trends® in Human–Computer Interaction*, *11*(2), 63–125. https://doi.org/10.1561/1100000015

Friedman, B., Hendry, D. G., Huldtgren, A., Jonker, C., Van den Hoven, J., & Van Wynsberghe, A. (2015). Charting the Next Decade for Value Sensitive Design. *Aarhus Series on Human Centered Computing*, *1*(1), 4. https://doi.org/10.7146/aahcc.v1i1.21619

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. (2003). Human values, ethics, and design. In J. A. Jacko & A. Sears (Eds.), *The Human-computer Interaction Handbook* (pp. 1177–1201). Hillsdale, NJ, USA: L. Erlbaum Associates Inc. Retrieved from http://dl.acm.org/citation.cfm?id=772072.772147

Friedman, B., & Kahn Jr, P. H. (2003). Human values, ethics, and design. *The Human-Computer Interaction Handbook*, 1177–1201.

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

Ghani, R. (2016). you say you want transparency and interpretability? Retrieved February 13, 2019, from https://dssg.uchicago.edu/2016/04/27/you-say-you-want-transparency-and-interpretability/

Gupta, A. (2017). Machine Learning Algorithms in Autonomous Driving. Retrieved July 16, 2019, from https://iiot-world.com/machine-learning/machine-learning-algorithms-in-autonomous-driving/

Hendrickson, C. S., & Van Nieuwstadt, M. J. (2018, December 6). System and method for platooning vehicles. Google Patents.

Johnson, A. M., & Axinn, S. (2013). The Morality of Autonomous Robots. *Journal of Military Ethics*, *12*(2), 129–141. https://doi.org/10.1080/15027570.2013.818399

Johri, A., & Nair, S. (2011). The role of design values in information system development for human benefit. *Information Technology & People*, *24*(3), 281–302. https://doi.org/10.1108/09593841111158383

Kahn Jr, P. H., Friedman, B., Freier, N., & Severson, R. (2003). Coding manual for children’s interactions with AIBO, the robotic dog–The preschool study. *University of Washington CSE Technical Report 03-04*, *3*.

Kamali, M., Dennis, L. A., McAree, O., Fisher, M., & Veres, S. M. (2017). Formal verification of autonomous vehicle platooning. *Science of Computer Programming*, *148*, 88–106. https://doi.org/https://doi.org/10.1016/j.scico.2017.05.006

Kamali, M., Linker, S., & Fisher, M. (2018). Modular Verification of Vehicle Platooning with Respect to Decisions, Space and Time. *ArXiv Preprint ArXiv:1804.06647*.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning* (pp. 2673–2682).

Klincewicz, M. (2015). Autonomous Weapons Systems, the Frame Problem and Computer Security. *Journal of Military Ethics*, *14*(2), 162–176. https://doi.org/10.1080/15027570.2015.1069013

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology*, *19*(2), 107–115.

Lee, S., & Son, Y.-J. (2008). Integrated human decision making model under belief-desire-intention framework for crowd simulation. In *Simulation Conference, 2008. WSC 2008. Winter* (pp. 886–894). IEEE.

Mecacci, G., & de Sio, F. S. (2019). Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics and Information Technology*, 1–13.

Mermet, B., & Simon, G. (2016). Formal verification of ethical properties in multiagent systems. In *CEUR Workshop Proceedings* (pp. 26–31).

Mordatch, I., & Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Mortier, H. H., Henderson, T., McAuely, D., & Crowcroft, J. (2014). Human-data interaction: the human face of the data-driven society. *Social Science Research Network*.

Oosterlaken, I. (2015). Applying Value Sensitive Design (VSD) to Wind Turbines and Wind Parks: An Exploration. *Science and Engineering Ethics*, *21*(2), 359–379. https://doi.org/10.1007/s11948-014-9536-x

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+of+Technological+Systems&hl=en&sa=X&ved=0ahUKEwjvtbqrkfHXAhUMMd8KHdXnDGEQ6AEIKDAA#v=onepage&q=The Social Construction of Technological Systems&f=false

Pitt, J., & Diaconescu, A. (2016). Interactive Self-Governance and Value-Sensitive Design for Self-Organising Socio-technical Systems. In *2016 IEEE 1st International Workshops on Foundations and Applications of Self\* Systems (FAS\*W)* (pp. 30–35). https://doi.org/10.1109/FAS-W.2016.20

Prokhorov, D. V. (2018, October 16). Mixed autonomous and manual control of autonomous vehicles. Google Patents.

Rao, A. S., & Georgeff, M. P. (1992). An abstract architecture for rational agents. In *3rd International Conference on Principles of Knowledge Representation and Reasoning* (pp. 439–449).

Roco, M. C. (2008). Possibilities for global governance of converging technologies. *Journal of Nanoparticle Research*, *10*(1), 11–29. https://doi.org/10.1007/s11051-007-9269-8

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

Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., … Vicente, R. (2017). Multiagent cooperation and competition with deep reinforcement learning. *PloS One*, *12*(4), e0172395.

Thornton, S. M., Lewis, F. E., Zhang, V., Kochenderfer, M. J., & Gerdes, J. C. (2018). Value sensitive design for autonomous vehicle motion planning. In *2018 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1157–1162). IEEE.

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

Umbrello, S. (2018). The moral psychology of value sensitive design: the methodological issues of moral intuitions for responsible innovation. *Journal of Responsible Innovation*, *5*(2), 186–200. https://doi.org/10.1080/23299460.2018.1457401

Umbrello, S. (2019a). Atomically Precise Manufacturing and Responsible Innovation. *International Journal of Technoethics*, *10*(2), 1–21. https://doi.org/10.4018/IJT.2019070101

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

Umbrello, S. (2020a). Combinatory and Complementary Practices of Values and Virtues in Design: A Reply to Reijers and Gordijn. *Filosofia*. Retrieved from https://www.academia.edu/43531194/Combinatory\_and\_Complementary\_Practices\_of\_Values\_and\_Virtues\_in\_Design\_A\_Reply\_to\_Reijers\_and\_Gordijn

Umbrello, S. (2020b). Meaningful Human Control over Smart Home Systems: A Value Sensitive Design Approach. *Humana.Mente Journal of Philosophical Studies*, *13*(37), 40–65.

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

Umbrello, S., Torres, P., & De Bellis, A. F. (2020). The future of war: could lethal autonomous weapons make conflict more ethical? *AI & SOCIETY*, *35*(1), 273–282. https://doi.org/10.1007/s00146-019-00879-x

van de Poel, I. (2013). Translating Values into Design Requirements. In D. P. Michelfelder, N. McCarthy, & D. E. Goldberg (Eds.), *Philosophy and Engineering: Reflections on Practice, Principles and Process* (pp. 253–266). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-7762-0\_20

van de Poel, I. (2017). Dealing with Moral Dilemmas through Design. In J. van den Hoven, S. Miller, & T. Pogge (Eds.), *Designing in Ethics* (pp. 57–77). Cambridge, UK: Cambridge University Press.

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

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

van Wynsberghe, A. (2012). *Designing Robots With Care: Creating an Ethical Framework for the Future Design and Implementation of Care Robots*. University of Twente. https://doi.org/10.3990/1.9789036533911

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

Vermaas, P. E., Hekkert, P., Manders-Huits, N., & Tromp, N. (2014). Design Methods in Design for Values. 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. 1–19). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6994-6\_10-1

Wachter, S., Mittelstadt, B., & Floridi, L. (2016). *European Union regulations on algorithmic decision-making and a “right to explanation.”* https://doi.org/10.1609/aimag.v38i3.2741

Wang, C., Gong, S., Zhou, A., Li, T., & Peeta, S. (2018). Cooperative adaptive cruise control for connected autonomous vehicles by factoring communication-related constraints. *ArXiv Preprint ArXiv:1807.07232*.

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

Woelfer, J. P., & Hendry, D. G. (2009). Stabilizing homeless young people with information and place. *Journal of the American Society for Information Science and Technology*, *60*(11), 2300–2312.

Wooldridge, M. (2002). *An Introduction to Multiagent Systems*. John Wiley & Sons.

Yampolskiy, R. V. (2017). What are the ultimate limits to computational techniques: verifier theory and unverifiability. *Physica Scripta*, *92*(9), 93001.

Yan, C., Xu, W., & Liu, J. (2016). Can you trust autonomous vehicles: Contactless attacks against sensors of self-driving vehicle. *DEF CON*, *24*.

Yoo, D. (2017). Stakeholder Tokens: a constructive method for value sensitive design stakeholder analysis. In *Proceedings of the 2017 ACM Conference Companion Publication on Designing Interactive Systems* (pp. 280–284). ACM.

1. Naturally some abstraction must be relied on with model-based programming since the real world cannot be fully captured by any model. However, that does not preclude that model cannot nor should not be continually improved, the very contrary is true. BDI verification tools for both system properties and continuous system controllers can take various forms to satisfy the hybrid architectures. Proposals for a hybrid between Gwendolen agent code and continuous control systems may prove to be a good way to envision this hybrid structure (Kamali et al., 2018). [↑](#footnote-ref-1)