International Journal of Social Robotics
https://doi.org/10.1007/s12369-021-00790-w

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

Steven Umbrello[1] · **Roman V. Yampolskiy**[2]

Accepted: 20 April 2021
© The Author(s) 2021

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. How human values are embodied and actualised in situ may ultimately prove to be harmful if not outright recalcitrant. For this reason, the values of stakeholders become of particular significance given the risks posed by opaque structures of intelligent agents. This paper explores how decision matrix algorithms, via the *belief-desire-intention* model for autonomous vehicles, can be designed to minimize the risks of opaque architectures. Primarily through an explicit orientation towards designing *for* the values of explainability and verifiability. In doing so, this research adopts the Value Sensitive Design (VSD) approach as a principled framework for the incorporation of such values within design. VSD is recognized as a potential starting point that offers a systematic way for engineering teams to formally incorporate existing technical solutions within 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 the changing ethical landscapes required for utilisation in autonomous vehicles.

Keywords Value sensitive design · Artificial intelligence · Autonomous vehicles · Explainability · Verifiability · Applied ethics

1 Introduction

The impacts and influences of autonomous systems, powered by artificial intelligence (AI), on society at large are no longer in question. Many of the ethical, social, and legal concerns, among others, derive from the black-boxed nature of the decision-making structures and logical pathways of autonomous systems. Opaque decision-making architectures do not permit designers or users to understand if their values have been substantively embodied. For this reason, the values of stakeholders become of particular significance given the risks posed by the opaque structures of intelligent agents (IAs). Autonomous vehicles (AVs) are one such AI-powered IA that often employs such potentially opaque decision-making structures and for this reason these have been taken as the object of analysis in this paper. In doing so, we propose a VSD approach as a principled framework for incorporating human values in design and thus retaining meaningful human control over them [1]. This works by applying VSD to formal verification (FV) policy for the decision matrix algorithms (DMAs) that can then be reasonably be employed in AV design.

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 both early on and throughout the subsequent design process [2]. 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 end-users among others [1, 3, 4].

To the best of our knowledge, this is the first paper to evaluate the suitability of the VSD approach to AI as it pertains to the values of explainability and verifiability for

✉ Steven Umbrello
steven.umbrello@unito.it

Roman V. Yampolskiy
roman.yampolskiy@louisville.edu

[1] Institute for Ethics and Emerging Technologies, University of Turin, Via Sette Comuni 45, Turin 10127, Italy

[2] Computer Engineering and Computer Science Department, University of Louisville, Louisville 40292, USA

Published online: 15 May 2021

AVs [5]. Prior literature on VSD has focused on its methodological foundations [6–8], its applicability to existing technologies, such as energy systems and care robotics [9–11], as well as its applicability to AI in general and other advanced technologies for instance, molecular manufacturing [12–15]. These studies provide useful information on the VSD approach in general, as well as why, and how it can be used for the development of IAs like AVs. However, none of these arguments focus specifically on the values of particular interest for the safe development of beneficial IAs. Similarly, unlike other research projects which focus on technologies as concepts, 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 during their development. 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 whilst Sect. 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 and Sect. 6 concludes this paper by summarizing its findings as well as by providing suggestions for potentially future research areas.

2 Emerging Issues with Autonomous Vehicles

Discussions on autonomous systems both in military and civil spheres continue to hold 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 provide a fertile ground for autonomous systems to proliferate across different domains with decreasing need for human command and control. Given the devolved power to these systems, the primary issue that emerges as a result of their introduction into society pertains to 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 that burden of responsibility lie? What if an AV kills a pedestrian on the sidewalk and the designers of that AV's programming cannot discern the decision pathway of the logic that resulted in 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 systems, particularly within the legal fields that are interested with legislation and the liability issues surrounding the governance of AV's which are already present on many roads [16–19].

Roads are continuously occupied by a variety of stakeholder groups such as vehicle drivers themselves, pedestrians, and cyclists. The implied values that are held by these stakeholders in large part governs their actions, reactions, and overall expectations to situations that can, and often do, arise on roads. The introduction and continual transition towards automation will most likely also be predicated by similar roadway values. Some of these existent roadway values are safety and adherence to 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 the design requirements 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 AV decision matrix algorithms (DMAs). Hence DMAs provide one such nexus where a VSD approach can intervene for explicit orientation to designing *for* human values in AVs.

DMAs are one of the primary algorithms used to form the machine learning set for AVs [20–22]. A popular example of a DMA is adaptive boosting (AdaBoosting), other employed algorithms include clustering processes such as K-means, pattern recognizers (classifiers), support vector machines, and regression formulas such as neural network regression. DMAs are chosen over the other implicated, and no less important algorithm systems because they function by methodically evaluating and ranking the efficacy of the relationship between data sets and values. The quality and fidelity of data sets is consequentially of high ethical importance, not only for the proper functioning of the system, but to ensure that the system does not display any unforeseen (or unforeseeable) emergent behavior on account of biased or other improper data categorisation. Nonetheless, because of their ability to assess each of the data-set elements for their relative significance, DMAs can be 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. This is attributed to the recognition and movement of environmental entities based on sensor input and the 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 to reduce errors in judgement. The training of these models with chosen data

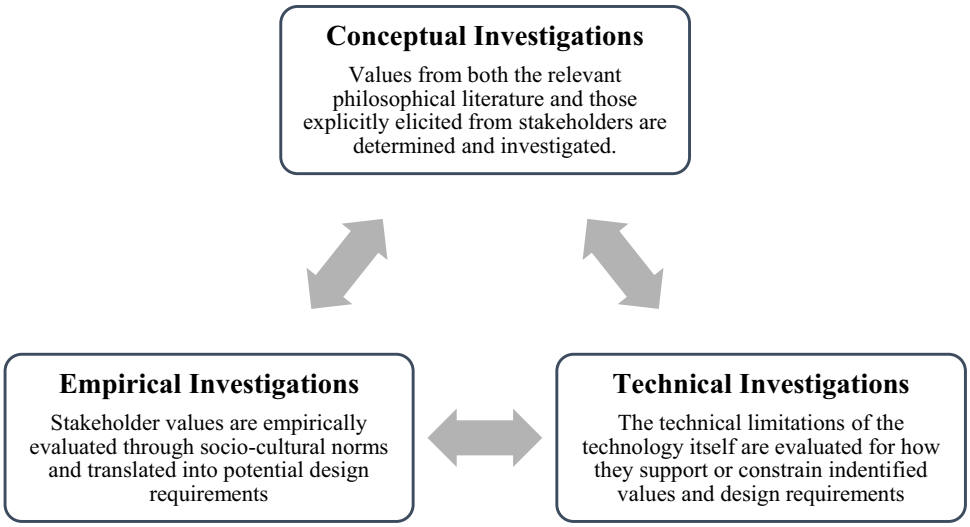

Fig. 1 The recursive VSD tripartite framework. *Source*: [28]

sets implicate values (i.e., which are chosen, also at the opportunity costs of chosen inputs) and provides the perfect place for designers to directly intervene with the intention of actively guiding the formation of these models through principled stakeholder engagement. This process helps reduce bias across the board from data and task setup, feature preprocessing to DMA model selection as a *by-design* interpretable model and recourse interface for deployment.

So why utilise VSD? The VSD approach is chosen because it is founded on the premise that technology value-laden and thus of significant ethical importance [23–25]. VSD is a principled approach to design that is divided into three distinct investigations (Fig. 1), referred to commonly as a 'tripartite methodology' consisting of conceptual investigations, empirical investigations, and technical investigations [7, 26]. The framework is designed to be iterative and recursively self-reflective as it aims to continually balance the values of both direct and indirect stakeholder groups throughout the design process. Because of this, the VSD methodology is typically employed in technological design where human values come into conflict with each other and where the design solutions are of considerable ethical concern [27]. Similar to how Friedman, Kahn, and Borning [7] used VSD in the realm of human–computer interaction to show how the values of privacy and usability needed to be balanced, Umbrello [14] demonstrated that stakeholders involved in the development of AI in the United Kingdom prized addition values. These comprised transparency, control, data privacy, and security, all of which need to be delicately balanced through careful stakeholder coordination and cooperation.

Nonetheless, the applicability of the VSD methodology to a variety of technological artefacts makes its acceptance by design groups more attractive because various scholars have demonstrated its ability to be easily adapted and streamlined into existing design practices. For example Timmermans et al. [15] embraced the VSD approach for the design of nanopharmaceuticals by adopting the values existent in the medical field and van Wynsberghe [29] similarly draws from the values of care to modify the VSD methodology for application in 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. Chen, Peng, and Grizzle [30] present a design for an obstacle avoidance algorithm for low-speed autonomous vehicles with the intent of balancing efficacy by reducing control effort while prioritizing safety through minimising pedestrian and obstacle collision. Similarly, Kamali et al. [31] propose formal verification (FV) of AV code, namely 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. However, the continual desire to balance the values of safety and efficacy leads to what VSD theorists call *moral overload*. This occurs when similar values are equally prized by enrolled stakeholders but are often at odds with each other when translated into technical design requirements [27, 32]. Similarly, aside from these two values, other values such as trust and control are implicated in AV design which can also come into conflict [16, 33–36]. 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 stakeholder expectation.

This paper proposes that VSD can not only help bridge the chasm in the design process of DMAs for AVs but a more comprehensive set of values can also be considered

that may be employed in AV design in addition to safety and effectiveness. These apply particularly to *explainability* and *verifiability* as well as proposals for how moral overload can be resolved through certain design flows. This investigation thus aims to demonstrate that an adapted form of the VSD approach for DMA model training can be applied with particular emphasis on obstacle avoidance models. In order for AVs to be deployed, their conclusions—a function of model training for DMA attribution—require FV to ensure traceable lines of decision-making logic that are receiver -contextualised and can thus be explained to/by designers to minimize incidents on the road. These also need to be repeatable and consistent in order to be verifiable and this FV policy is designed using VSD as a framework. This *value-sensitively designed* DMA, although limited to the AVs controller-code and not applied to the full model of the autonomous system, nonetheless captures the decision-making structure of the agent code which 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, technical limits and constraints.

3 Value Sensitive Design

The VSD methodology is traditionally conceived of as a tripartite methodology consisting of three stages of investigations: conceptual, empirical and technical [8] (see Fig. 1). The first of these, conceptual investigation, consists of answering the following questions: Who are the stakeholders? What are the values related to the technology in question? Where do certain parameters begin and end when discussing the bounds of usability versus conflicting values such as transparency and privacy or safety and efficacy? Who are the direct versus indirect stakeholders? When are the agreed 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 scope of conceptual investigation [37].

The second phase, empirical investigation, aims to use both quantitative and qualitative analyses to determine if the distilled conceptual values can meet the need of stakeholders in design. This includes 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 [12]. This stage ultimately aims to determine if the design of a technology maps onto the conceptual results and if not, a recursive feedback to conceptual investigations is the required to determine how those values can be better mapped onto the design.

Finally, technical investigation looks at the 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 minimising unwanted or potentially emergent problems. The technical questions become important in the application of identified values given that they can constrain how they are instantiated in the design [38].

Together these three processes are meant to be iterative, feeding into one another until alignment between them becomes harmonized. Designers tend to already engage in self-feedback and redesign until they meet their desired criteria. Through prototyping and small-scale deployment designers can determine if and how AVs can manifest emergent and unforeseen values. In such cases, unwanted emergent values trigger further iterations of the VSD process. The VSD methodology enables a more principled way of formalizing this otherwise implicit practice and better ensure value-alignment both in the early phases and throughout the subsequent design process [39, 40].

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

In beginning the VSD process, one of the most critical steps involved is identifying direct and indirect stakeholders. 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 the 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. The relationship between direct and indirect stakeholders is dynamic and 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 [41], stakeholder tokens [42] and Envisioning Cards [43].

When taking stakeholders into 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 the design [44, 45]. What becomes important is to highlight which values are

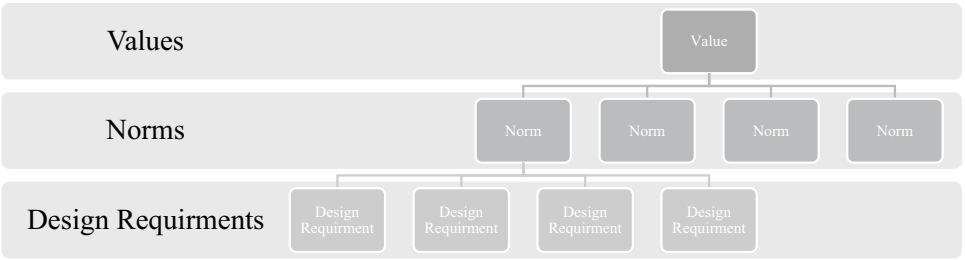

Fig. 2 Values hierarchy. *Source*: [49]

implicated and how desired values can be achieved, particularly amongst other aspirations that may be in tension (i.e., privacy and security) [27]. Tools such as value source analysis [46], the value-oriented coding manual [47] and/or value-oriented mock-ups or prototypes [48] 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 the stakeholders involved further values can be identified such as *trust, autonomy, transparency*, and *privacy* [i.e., 14]. Because each of these values can lead to different design requirements and flows, it becomes important to ensure they are effectively conceptualized and balanced to enable translation into engineering goals. Van de Poel [49] uses a hierarchy to help designers translate value through norms into tangible design requirements (Fig. 2). This research uses the value of transparency as an example 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.

Transparency is often cited as a desirable commodity in much of the AI literature [50–52]. However, the term is sensitive to contextualize, and is far more nuanced than is often considered. Although often beneficial within the context of algorithmic verifiability and understanding, there are instances where transparency may not be beneficial for stakeholders. These can occur when too much information, such as that of data subjects, becomes overly accessible to a wide variety of potential users consequently denigrating certain privacy norms. It is important to remember that just because a system is transparent, this does not necessarily mean that any given user can understand it (i.e., the need for receiver-contextualised explainability) nor does the accessibility of user data, such as driving habits, mean that the system has properly satisfied our transparency values. The efficacy of transparency is similarly context dependent on goals and definitions. It can take the form of designers being able to determine how well a system is preforming and how it can be improved. This helps the public to understand the strengths and drawbacks of a particular system and develops trust. It also allows users and designers 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 [53]. This is not an exhaustive list of ways to conceptualize transparency but the examples listed all demonstrate that transparency is construed as a general benefit to 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 ensuring that no information is omitted that may be crucial to the agents implicated.

Nevertheless, transparency can similarly come into tension with other important values, particularly when considering AVs. The issues arise 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 that requiring the source code of systems be fully open, can not only lead to the manipulation of such code, but can also disincentivize industry leaders to innovate due to the lack of proprietorship [54, 55].

However, perhaps the most obvious tension that the value of transparency can have is when compared with 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 often experience tension where stakeholders want both a right to data privacy but also transparency over how such systems function. The tension is most obvious where greater transparency can lead to increased trust in a system (where the interests between the system and the stakeholder align) but also where privacy of a system also encourages stakeholder 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) 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 consider it as a means of supporting or constraining other important values as a design flow. In summary, transparency is an instrumental value, a value-in-process is how it should be conceptualized rather than as an end-value [56].

One of the initial ways of conceptualizing these design flows through VSD, and encouraging engineers to accept and adopt the methodology, is through integration with similar practices and theories and in this instance, 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 and verified in greater depth [31, 57]. 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 *why* it chooses to do so [58]. This rational agent approach not only provides transparency, it also promotes self-improving design flows which in turn promotes its acceptance by engineering teams and industry.

Hence, the models from which DMA attribution can begin to be conceptualized may originate within 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 [59–62]. 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 the 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. On such advantage 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 demarcating scenario selection and scenario execution [63]. Such a 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]

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 through platooning research, whereby a 'platoon' or convoy of AVs synchronically follow a lead vehicle that is under human control [64, 65]. 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 [66].

The following discussion outlines some perceived implications of this approach as well as limitations and potential further research avenues that may prove beneficial when applied to full autonomous vehicles.

5 Discussion

There are numerous drawbacks and as such, many fruitful areas of potential future research that this paper identifies. Firstly, the rigid structure of the BDI hierarchy and DMA control system in general naturally precludes any built-in learning, planning, and adaptation from environmental inputs or past events. What this does 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. This paper proposes an explicit modeling structure that promotes transparency as a means towards enhancing safety and operability of AVs, rather than an end in itself. Similarly, this can build trust amongst the designers and users in their ability to understand the rational decisions taken by the agent within a certain set of input parameters in the models. Further research should concentrated upon how the VSD methodology can balance the value requirements distilled by design teams while considering advanced machine learning systems. Indeed, there are many ways of modelling and implementing discrete decision-making joint to continuous control mechanisms but remains a challenge to design and implement in many application areas. In addition, this paper promotes the importance of the decision matrix algorithms which are already well known in the community and of which there are both technical and methodological challenges concerning their design and implementation. Even without considering concepts such as explainability. Whether this is even feasible is not the subject of this research, 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

[1] Naturally some abstraction must be relied on with model-based programming since the real world cannot be fully captured within any such model. However, that does not preclude that the model cannot, nor should not, be continually improved as 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 an efficient way to envision this hybrid structure [64].

design flow. This can then be used to guide engineers and programmers in conceptualizing other important values where transparency can be utilised as a way to either support or 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 *humansystem* 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 [1, 66]. 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 [67–69].

Another potential avenue for further research would be in the transparency and interpretability dynamics of machine-machine relationships. Steps in this direction have already been taken to consider how autonomous systems can communicate, coordinate and execute tasks together [70–72]. The organization and dynamics of these multi-agent ensembles should be further explored for a number of reasons. The first would be in the cooperation between differing systems to autonomously institute extensible concepts and go beyond the 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 humans 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 focus their investigation is upon methods 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 in 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 in isolation which comes with a host of associated issues [73, 74]. 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.

6 Conclusions

In this proposal for 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 in which designers can formally connect human values to design requirements. This paper proposes that the decision matrix algorithms of AVs provide a potentially fruitful starting point for considering 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. Consequently, the VSD methodology directly enrolls not only members of the public and industry representatives, 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 both the early phases and throughout the design process so that policy and technology can harmoniously align.

Acknowledgements We would like to thank the two anonymous reviewers whose comments have been useful in revising the original manuscript. Any remaining errors are the authors alone. The views in the paper are the authors alone and not the views of the Institute for Ethics and Emerging Technologies.

Funding Open access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Santoni de Sio F, van den Hoven J (2018) Meaningful human control over autonomous systems: a philosophical account. Front Robot AI. 5:15
2. Friedman B (1997) Human values and the design of computer technology. In: Friedman B (ed) CSLI Publications
3. van den Hoven J, Manders-Huits N (2009) Value-sensitive design. In: A companion to the philosophy of technology, Wiley-Blackwell, pp 477–480. https://doi.org/10.1002/9781444310795.ch86
4. Friedman B, Kahn Jr PH (2003) Human values, ethics, and design. Hum-Comput Interact Handb:1177–1201
5. Yampolskiy RV (2017) What are the ultimate limits to computational techniques: verifier theory and unverifiability. Phys Scr 92(9):93001
6. Umbrello S (2018) The moral psychology of value sensitive design: the methodological issues of moral intuitions for responsible innovation. J Respons Innov 5(2):186–200. https://doi.org/10.1080/23299460.2018.1457401
7. Friedman B, Kahn PH, Borning A, Huldtgren A (2013) Value sensitive design and information systems. In: Doorn N, Schuurbiers D, van de Poel I, Gorman ME (eds) Early engagement and new technologies: opening up the laboratory. Springer, Dordrecht, pp 55–95. https://doi.org/10.1007/978-94-007-7844-3_4
8. Friedman B, Hendry DG, Borning A (2017) A survey of value sensitive design methods. Found Trends Hum-Comput Interact 11(2):63–125. https://doi.org/10.1561/1100000015
9. Oosterlaken I (2015) Applying value sensitive design (VSD) to wind turbines and wind parks: an exploration. Sci Eng Ethics 21(2):359–379. https://doi.org/10.1007/s11948-014-9536-x
10. 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
11. van Wynsberghe A (2016) Service robots, care ethics, and design. Ethics Inf Technol 18(4):311–321. https://doi.org/10.1007/s10676-016-9409-x
12. Umbrello S, De Bellis AF (2018) A value-sensitive design approach to intelligent agents. In: Yampolskiy RV (ed) Artificial intelligence safety and security, CRC Press, pp 395–410. https://doi.org/10.13140/RG.2.2.17162.77762
13. Umbrello S (2019) Beneficial artificial intelligence coordination by means of a value sensitive design approach. Big Data Cogn Comput 3(1):5. https://doi.org/10.3390/bdcc3010005
14. Umbrello S (2019) Atomically precise manufacturing and responsible innovation. Int J Technoethics 10(2):1–21. https://doi.org/10.4018/IJT.2019070101
15. 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
16. Contissa G, Lagioia F, Sartor G (2017) The ethical knob: ethically-customisable automated vehicles and the law. Artif Intell Law 25(3):365–378. https://doi.org/10.1007/s10506-017-9211-z
17. Thornton SM, Lewis FE, Zhang V, Kochenderfer MJ, Gerdes JC (2018) Value sensitive design for autonomous vehicle motion planning. In: 2018 IEEE intelligent vehicles symposium (IV); IEEE, pp 1157–1162
18. 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
19. Contissa G, Lagioia F, Sartor G (2017) Accidents involving autonomous vehicles: legal issues and ethical dilemmas. JUSLETTER:1–7
20. Gupta A (2019) Machine learning algorithms in autonomous driving https://iiot-world.com/machine-learning/machine-learning-algorithms-in-autonomous-driving/. Accessed 16 Jul 2019
21. 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
22. Leben D (2017) A Rawlsian algorithm for autonomous vehicles. Ethics Inf Technol 19(2):107–115
23. 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
24. Davis J, Nathan LP (2014) Value sensitive design: applications, adaptations, and critiques. In: van den Hoven J, Vermaas PE, van de Poel I (eds) Handbook of ethics, values, and technological design: sources, theory, values and application domains. Springer, Dordrecht, pp 1–26. https://doi.org/10.1007/978-94-007-6994-6_3-1
25. Friedman B, Hendry DG, Huldtgren A, Jonker C, Van den Hoven J, Van Wynsberghe A (2015) Charting the Next decade for value sensitive design. Aarhus Ser Hum Centered Comput 1(1):4. https://doi.org/10.7146/aahcc.v1i1.21619
26. Friedman B, Kahn PH Jr (2003) Human values, ethics, and design. In: Jacko JA, Sears A (eds) The human-computer interaction handbook. L. Erlbaum Associates Inc., Hillsdale, pp 1177–1201
27. van den Hoven J, Lokhorst GJ, van de Poel I (2012) Engineering and the problem of moral overload. Sci Eng Ethics 18(1):143–155. https://doi.org/10.1007/s11948-011-9277-z
28. Umbrello S (2020) Meaningful human control over smart home systems: a value sensitive design approach. Humana Mente J Philos Stud 13(37):40–65
29. van Wynsberghe A (2013) Designing robots for care: care centered value-sensitive design. Sci Eng Ethics 19(2):407–433. https://doi.org/10.1007/s11948-011-9343-6
30. Chen Y, Peng H, Grizzle J (2018) Obstacle avoidance for low-speed autonomous vehicles with barrier function. IEEE Trans Control Syst Technol 26(1):194–206. https://doi.org/10.1109/TCST.2017.2654063
31. Kamali M, Dennis LA, McAree O, Fisher M, Veres SM (2017) Formal verification of autonomous vehicle platooning. Sci Comput Program 148:88–106. https://doi.org/10.1016/j.scico.2017.05.006
32. van de Poel I (2017) Dealing with moral dilemmas through design. In: van den Hoven J, Miller S, Pogge T (eds) Designing in ethics. Cambridge University Press, Cambridge, pp 57–77
33. Abraham H, Lee C, Brady S, Fitzgerald C, Mehler B, Reimer B, Coughlin JF (2017) Autonomous vehicles, trust, and driving alternatives: a survey of consumer preferences. In: Transportation research board 96th annual meeting, Washington, pp 8–12
34. Yan C, Xu W, Liu J (2016) Can you trust autonomous vehicles: contactless attacks against sensors of self-driving vehicle. DEF CON 24
35. Prokhorov DV (2018) Mixed autonomous and manual control of autonomous vehicles. Google Patents
36. Wang C, Gong S, Zhou A, Li T, Peeta S (2018) Cooperative adaptive cruise control for connected autonomous vehicles by factoring communication-related constraints. http://arxiv.org/abs/1807.07232
37. Denning T, Kohno T, Levy HM (2013) A framework for evaluating security risks associated with technologies used at home. Commun ACM. https://doi.org/10.1145/2398356.2398377
38. Friedman B, Kahn Jr PH (2002) Value sensitive design: theory and methods. Univ Washingt Tech https://doi.org/10.1016/j.neuropharm.2007.08.009

39. Friedman B, Hendry DG (2019) Value sensitive design: shaping technology with moral imagination. Mit Press, Cambridge
40. Umbrello S (2020) Combinatory and complementary practices of values and virtues in design: a reply to Reijers and Gordijn. Filosofia
41. 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
42. 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
43. Friedman B, Hendry DG (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
44. Winner L (2003) Do artifacts have politics? Technol Futur 109(1):148–164. https://doi.org/10.2307/20024652
45. Pinch T, Bijker WE (1987) The social construction of facts and artifacts. In: Bijker W E, Hughes TP, Pinch T (eds) The Social construction of technological systems : new directions in the sociology and history of technology, MIT Press, p 405
46. 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, Springer, pp 449–468
47. Kahn PH Jr, Friedman B, Freier N, Severson R (2003) Coding manual for children's interactions with AIBO, the robotic dog-the preschool study. Univ Washingt CSE Tech Rep 03–04:3
48. Woelfer JP, Hendry DG (2009) Stabilizing homeless young people with information and place. J Am Soc Inf Sci Technol 60(11):2300–2312
49. van de Poel I (2013) Translating values into design requirements. In: Michelfelder DP, McCarthy N, Goldberg DE (eds) Philosophy and engineering: reflections on practice, principles and process. Springer, Dordrecht, pp 253–266. https://doi.org/10.1007/978-94-007-7762-0_20
50. Mortier HH, Henderson T, McAuely D, Crowcroft J (2014) Human-data interaction: the human face of the data-driven society. Soc Sci Res Netw
51. Johri A, Nair S (2011) The role of design values in information system development for human benefit. Inf Technol People 24(3):281–302. https://doi.org/10.1108/09593841111158383
52. Vermaas PE, Hekkert P, Manders-Huits N, Tromp N (2014) Design methods in design for values. In: van den Hoven J, Vermaas PE, van de Poel I (eds) Handbook of ethics, values, and technological design: sources, theory, values and application domains. Springer, Dordrecht, pp 1–19. https://doi.org/10.1007/978-94-007-6994-6_10-1
53. Mecacci G, de Sio FS (2019) Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. Ethics Inf Technol:1–13
54. Ghani R (2019) you say you want transparency and interpretability? https://dssg.uchicago.edu/2016/04/27/you-say-you-want-transparency-and-interpretability/. Accessed 13 Feb 2019
55. Roco MC (2008) Possibilities for global governance of converging technologies. J Nanoparticle Res 10(1):11–29. https://doi.org/10.1007/s11051-007-9269-8
56. Boscoe B (2019) Creating transparency in algorithmic processes. Delphi Interdiscip Rev Emerg Technol 2 (1)
57. Wooldridge M (2002) An introduction to multiagent systems, John Wiley & Sons
58. Fisher M, Dennis L, Webster M (2013) Verifying autonomous systems. Commun ACM 56(9):84–93
59. Rao AS, Georgeff MP (1992) An abstract architecture for rational agents. In: 3rd international conference on principles of knowledge representation and reasoning, pp 439–449
60. Cointe N, Bonnet G, Boissier O (2016) Multi-agent based ethical asset management. In: CEUR workshop proceedings, pp 52–57
61. Caillou P, Gaudou B, Grignard A, Truong CQ, Taillandier PA (2017) Simple-to-use BDI architecture for agent-based modeling and simulation. In: Advances in social simulation 2015, Springer, pp 15–28
62. 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; IEEE, pp 886–894
63. Dennis LA, Fisher M, Lincoln NK, Lisitsa A, Veres SM (2016) Practical verification of decision-making in agent-based autonomous systems. Autom Softw Eng 23(3):305–359
64. Kamali M, Linker S, Fisher M (2018) Modular verification of vehicle platooning with respect to decisions, space and time. http://arxiv.org/abs/1804.06647
65. Hendrickson CS, Van Nieuwstadt MJ (2018) System and method for platooning vehicles. Google Patents
66. Calvert SC, Mecacci G, Heikoop DD, de Sio FS (2018) Full platoon control in truck platooning: a meaningful human control perspective. In: 2018 21st international conference on intelligent transportation systems (ITSC); IEEE, pp 3320–3326
67. Johnson AM, Axinn S (2013) The morality of autonomous robots. J Mil Ethics 12(2):129–141. https://doi.org/10.1080/15027570.2013.818399
68. Klincewicz M (2015) Autonomous weapons systems, the frame problem and computer security. J Mil Ethics 14(2):162–176. https://doi.org/10.1080/15027570.2015.1069013
69. Umbrello S, Torres P, De Bellis AF (2020) The future of war: Could lethal autonomous weapons make conflict more ethical? AI Soc 35(1):273–282. https://doi.org/10.1007/s00146-019-00879-x
70. Mordatch I, Abbeel P (2018) Emergence of grounded compositional language in multi-agent populations. In: Thirty-second AAAI conference on artificial intelligence
71. Tampuu A, Matiisen T, Kodelja D, Kuzovkin I, Korjus K, Aru J, Aru J, Vicente R (2017) Multiagent cooperation and competition with deep reinforcement learning. PLoS ONE 12(4):e0172395
72. Mermet B, Simon G (2016) Formal verification of ethical properties in multiagent systems. In: CEUR workshop proceedings. pp 26–31
73. Bau D, Zhou B, Khosla A, Oliva A, Torralba A (2017) Network dissection: quantifying interpretability of deep visual representations. http://arxiv.org/abs/1704.05796
74. 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

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Steven Umbrello currently serves as the Managing Director at the Institute for Ethics and Emerging Technologies. He is also currently an expert ethics consultant at Ethical Intelligence Associates Ltd. where he provides consultancy for companies to gain a competitive edge by mitigating ethical risks and formulating custom tools and protocols to embed ethics into their practices that align with current ethical guidelines and company values. Currently, his main area of research revolves around Value Sensitive Design (VSD), its philosophical foundations as well as its potential application to emerging technologies such as artificial intelligence and Industry 4.0.

Roman V. Yampolskiy is a Tenured Associate Professor in the department of Computer Engineering and Computer Science at the Speed School of Engineering, University of Louisville. He is the founding and current director of the Cyber Security Lab and an author of many books including Artificial Superintelligence: a Futuristic Approach. During his tenure at UofL, Dr. Yampolskiy has been recognized as: Distinguished Teaching Professor, Professor of the Year, Faculty Favorite, Top 4 Faculty, Leader in Engineering Education, Top 10 of Online College Professor of the Year, and Outstanding Early Career in Education award winner among many other honors and distinctions. Yampolskiy is a Senior member of IEEE and AGI; Member of Kentucky Academy of Science, and Research Advisor for MIRI and Associate of GCRI.