Ethics and Information Technology (2023) 25:11
https://doi.org/10.1007/s10676-023-09687-w

ORIGINAL PAPER

Value Sensitive Design for autonomous weapon systems – a primer

Christine Boshuijzen-van Burken[1]

Published online: 11 February 2023

Abstract
Value Sensitive Design (VSD) is a design methodology developed by Batya Friedman and Peter Kahn (2003) that brings in moral deliberations in an early stage of a design process. It assumes that neither technology itself is value neutral, nor shifts the value-ladennes to the sole usage of technology. This paper adds to emerging literature onVSD for autonomous weapons systems development and discusses extant literature on values in autonomous systems development in general and in autonomous weapons development in particular. I identify opportunities, such as public debates, and threats, such as the classified nature of the design process, for VSD in autonomous weapons development. This paper contributes to academic debates about the non-neutrality of technology by pointing out that values have been and can be explicitly designed into autonomous systems. It is informative for policy makers and designers who are tasked with developing actual autonomous weapons or policies around such systems, as they learn about an established design methodology that is sensitive to societal concerns and industry needs and that can be applied to autonomous weapons systems.

Keywords Value Sensitive Design · Autonomous Weapons · Autonomous Systems · Ethics of Technology · Non-neutrality of Technology

Introduction

What is the relationship between values and autonomous weapon systems, if any? Miller points out that there is a relationship between guns and values, by referring to the well-known slogan "Guns don't kill people, people kill people" in his discussion of the Value-Neutrality Thesis (VNT) (2020). According to the VNT, technology is neither good nor bad and only its uses have moral or other value, not the technology itself. In other words, guns are neither good nor bad (i.e. guns are value neutral), but the way in which the gun is used can be good or bad (i.e. its usage reflects a value – or lack thereof). Both the 'guns don't kill people' slogan and the VNT become increasingly more relevant in an era in which autonomous systems are becoming part of our everyday life, including the potential for autonomous systems that are put in place for our defence and that can employ lethal force. What if, at some point, guns kill people without a human directly intervening to make the decision of whom and when to kill? Can we then hold on to the claim that this technology, namely an autonomous weapon, including its actions (the act of killing a person), is value neutral? With such scenarios becoming more plausible, academics, layman, politicians and lawmakers alike are debating these issues from different angles. Some argue for a total ban on autonomous weapons, with many of said opponents often referring to autonomous weapons as 'killer robots' (e.g. the campaign to stopkillerrobots.org), while others argue that we may have a moral obligation to develop and use autonomous weapons (e.g. Strawser 2010; Arkin, 2010). In this article I defend a view that neither sees technology as value neutral, nor shifts the value-ladennes to the sole usage of technology. I propose deploying a design methodology for autonomous systems that does not merely focus on functional, economical or strategic performance, but in which concerns of an ethical and legal nature are at the heart of their development. I approach the issue of the development of autonomous weapons systems from a value sensitive design (VSD) perspective, initially developed by Batya Friedman and Peter Kahn in the early 1990s. They argued for bringing in moral deliberations into computer systems design through a conceptual analysis of human

✉ Christine Boshuijzen-van Burken
christinevanburken@gmail.com

[1] University of New South Wales (UNSW), Canberra, Australia

agency (2003). In later publications Friedman et al. provide additional values that are important for the design of information systems (human welfare, ownership and property, privacy, freedom from bias, universal usability, trust, autonomy, informed consent, accountability, courtesy, identity, calmness, and environmental sustainability). Others used VSD as a method to include values such as inclusivity and diversity in surveillance technology by Briggs and Thomas (2015), Van Wynsberghe (2013) proposed the design robots around the value of care, Jenkins et al. (2020) for the value of justice in design of energy systems, van de Kaa et al. (2020) for privacy, environmental sustainability, compatibility, trust, reliability, cost-effectiveness, and justice. Nickel (2015) promotes to design for trust and Boyed (2022) designed a tool for detecting fairness, transparency, privacy and accountability issues in Machine Learning. In a similar vein I propose to bring in moral deliberations into an early stage of autonomous weapons design. I recognize with Friedman and colleagues that the design space for technological innovation encompasses not only the technical design space, but also the corresponding socio-structural one (2017). This is particularly true for Defence and security related technology, as military and security technology development and acquisition is often highly regulated and moreover, their use ultimately rests on a delegated mandate from states and their societies. The status of this article can therefore be viewed as a primer that prepares for the integration of value sensitive design into the design and acquisition processes for autonomous systems for defence. Responsible development of autonomous systems in Defence encompasses both social structures and technical design, as I have argued elsewhere for military technology in general (Boshuijzen-van Burken, 2016; Boshuijzen-van Burken & Van Bezooijen, 2015; Boshuijzen-van Burken, 2021). In this article I discuss key authors that have suggested VSD for autonomous systems design in Defence (e.g. Santoni de Sio & van den Hoven, 2018; Umbrello 2019; Umbrello et al., 2020; Verdiesen, 2017; Verdiesen et al., 2019; Verdiesen & Dignum, 2022), and generate weaknesses, strengths and further questions for the VSD approach for autonomous systems in Defence. I furthermore connect the VSD discussion to the VNT discussion and make an appeal to the inclusion of societal values by referring to a special clause relevant to international humanitarian law, called the Martens Clause.

I begin the article by laying bare some assumptions that underlie my research on VSD for autonomous weapons systems. I then explain the VSD methodology, elaborate on literature in which, first, values and autonomous systems in general, and later, values and autonomous weapons in particular are discussed, and I provide an overview of the opportunities and threats for VSD in autonomous weapons development. I conclude with a general reflection on VSD for autonomous weapons and suggestions for further research.

Assumptions and definitions

Two important assumptions underlie this article. One, I assume that the use of weapons, including autonomous weapons irrespective of how we define them, is not an evil thing in itself, so long as they are used to promote justice and are used within the legal boundaries of a state, which has the mandate on the use of force.[1] I am fully aware that this assumption may be contested, but that is not the focus of this article, although there are points of contact between the assumptions on the legitimate use of force through autonomous weapons and the way in which we design autonomous weapons systems. The focus of this article is proactively focused on a design methodology for autonomous weapons, so that they can promote justice. Weapons may be developed based on other values, such as for establishing security and imposing peace, which are legitimate alternative candidates. However, peace and security can exist in the absence of justice, for example, through dictatorial regimes where the use of force is employed to suppress resistance to the 'peaceful' order that is upheld by an unjust state (Wolterstorff, 1983). I am furthermore aware that justice is not a univocal term in academic literature, but for the sake of this paper I appeal to an intuitive understanding of justice and note that Friedman and Hendry (2019) argue that justice is one of the three objective moral values in all VSD projects (see also, Umbrello 2020).

Two, it is assumed that despite epistemological differences and practical difficulties that are inevitable when connecting reflective and empirical practices, such as ethics and engineering, it is possible to improve technological artefacts through incorporating moral deliberations in the design process. I find VSD a promising approach to this challenge and this stems in part from an acceptance of VSD's central premise, that technology is not value-neutral, and this includes socio-technical dynamics that influence design processes. In other words, technology is implicitly or explicitly laden with values that inevitably originate from – and are of ethical importance to – individuals and society.

We now continue with definitions of autonomous systems, of which several exist. For example, Floridi and

[1] I refer to Dooyeweerd (1953) for an in depth argument regarding the use of force in relation to the state: "In whatever way we consider the matter, this foundational function of the genotype "State" can nowhere else be found but in an *internal monopolistic organization of the power of the sword over a particular cultural area within territorial boundaries.* The reader should remember that this typical historical structural function may in no way be naturalistically be misinterpreted. [.] it is a normative structural function […] which can be realized in a better or worse way." (1953, Vol III, 413).

Sanders mean with "autonomy" that the system "can change state without stimulus" (2004, p. 357). The Assuring Autonomy Body of Knowledge by the Assuring Autonomy International Programme at York University uses the following concise definition of autonomy: "the capability to make decisions free from human control." (Assuring Body of Knowledge, n.d.) In a further elaboration on autonomy, they state that they mean technology that has decision-making capability and authority. Part of a discussion on autonomous systems relates to degrees of autonomy that one wishes to allocate to an autonomous system. This observation may be important for the VSD of autonomous weapons, from a procedural and methodological perspective. It is procedurally relevant, because if in the definitional and terminological phase of system development certain user scenarios are excluded, simply because they do not match the criteria of "autonomous system", certain prospective user groups may not be consulted as they are no longer deemed stakeholders. It is methodologically relevant for settling conflicting values in even this earliest possible definitional and terminological phase of system development and potentially impinge upon which tasks can or should be allocated to an autonomous system in theory and practice. For example, how stakeholders interpret the value of systems reliability may greatly differ depending on if a task is allocated to an autonomously operating element in the system or if that task is allocated to a human operator (consider the tasks of navigation, or target selection, or target engagement).

A definition of autonomous weapons is difficult to provide, since this is an emerging technology and there is to date no internationally agreed upon definition (Advisory council on International Affairs, 2015; GGE LAW, 2019). In a recent overview by Taddeo and Blanchard (2022) several options were discussed. For the sake of this article I propose the definition from the International Committee of the Red Cross: "weapon systems with autonomy in their "critical functions" of selecting and attacking targets" (ICRC 2019, 2).

Lastly, I provide a (rather loose) definition of values which I deem suitable for the purpose of this article. There has been much disagreement in the history of philosophy over whether values exist, in what way and if and how they are different from facts. Friedman, Kahn and Borning propose a relatively loose definition of values in relationship to VSD. They state that a value is something that "[…] a person or group of people consider important in life" (Friedman et al., 2006, p. 349).

Value Sensitive Design

VSD was put forward by Batya Friedman and others as a conceptual tool to deliberately incorporate values into technological and socio-technical design (Friedman, 1996; Friedman et al., 2002, 2006; Friedman & Kahn, 2003; Umbrello & van de Poel 2021; van den Hoven et al., 2015; Winkler & Spiekermann, 2021). The aim of VSD is to include moral values in design; for example, human welfare, ownership and property, privacy, freedom from bias, universal usability, trust, autonomy, informed consent, accountability, identity, calmness, environmental sustainability, responsibility, safety, freedom (Friedman & Kahn, 2003; van den Hoven et al., 2015) in a technical design process. VSD sets out an integrative and iterative tripartite methodology, juggling and keeping in play the results of conceptual, empirical, and technological investigations (van de Poel & Royakkers, 2011).

During the conceptual phase of design, different questions regarding the notion of values are addressed, such as what counts as a value, whose values should be included, how value trade-offs should be dealt with, weather some values have greater weight, or if there are values that trump all other values in the technological design. (Friedman et al., 2002). Friedman and colleagues distinguish between two classes of stakeholders: direct and indirect. With direct stakeholders they mean parties – individuals or organizations – who interact directly with the technology. With indirect stakeholders they refer to all other parties who are affected by the use of the technology and who are often ignored in the design process.

Empirical investigations concern the human context within which the technical artifact will be situated and encompasses any human activity that can be observed, measured, or documented (Friedman et al., 2002). Questions that can be asked during the empirical phase are about stakeholders' individual values and how they prioritize competing values in design trade-offs, including usability considerations. Friedman and colleagues furthermore identify four relationships between usability and human values with ethical import:

> "First, a design can be good for usability and independently good for human values with ethical import (e.g., a highly usable adaptable interface can also promote user autonomy). Second, a design can be good for usability but at the expense of human values with ethical import (e.g., a highly usable system for surveillance that undermines the value of privacy). Third, a design can be good for human values with ethical import but at the expense of usability (e.g., a web browser setting that asks the user to accept or decline each cookie individually supports the value of informed consent, but is largely unusable due to the nuisance factor). And fourth, a design good for usability may be necessary to support human values with

> ethical import (e.g., in order to have a fair national election using a computerized voting system, all citizens of voting age must be able to use the system)." (Friedman et al., 2002, p. 3).

At this point I flag the importance of being aware of these complexities in ethical design and point out that on some occasions it may be necessary to give ground judiciously on one or the other to create a viable design. This challenge has been referred to as "moral overload" by van de Hoven et al. and is considered positively rather than negatively by these authors, as it can function as a driver for innovation in engineering in design (2012).

The technical investigations are where the focus shifts from the people or organizations that hold values, to the technology that either embodies, supports, or hinders human values in case of existing technologies, or how values identified in the conceptual investigation can be incorporated during the proactive design of technologies.

Numerous methods for implementing VSD in practice have been developed or adapted from existing methods in the social sciences, human-computer interaction, security, and other disciplines. These methods include semi-structured interviews, focus groups, and ethnography (see Friedman & Hendry (2019) for an overview of 17 VSD methods found in literature, along with strategies and heuristics for skilful practice[2]).

As the previous sections have made clear, values can be designed explicitly or implicitly into a technology and they can be applied or identified at different levels of a technology. The conceptualization and ideation of designing artefacts can itself stem from a certain value and this can be referred to as design for values (Vermaas et al., 2015). For example, video conference software Zoom allows for virtual backgrounds, so that others in the online meeting room will only see the face of the online collaborator against an artificial background. The actual or real background (i.e. the room or office where the person who is in the meeting sits) becomes invisible and is replaced by a picture that is chosen by the Zoom user. For some users the background option provides a great way to hide indications that may reveal someone's socio-economic background, and therefore the background functionality of Zoom embodies the value of equity amongst Zoom users. However, the actual usage of the virtual background option is only available to a certain group of people, namely those with lighter skin complexion. Darker-skinned people reported difficulties using the option (Costley, 2020), because the threshold and sensitivity for user versus background contrast is designed in such a way that the virtual background option does not detect enough contrast between a dark-skinned face and its background. Because the software cannot detect sufficient contrast it hides the user completely behind the virtual background, or only shows the eyes or teeth of the user against the virtual background. In the case of Zoom backgrounds, the value of equity was not realized, despite the best intentions of the designers. The Zoom example demonstrates where values are explicitly (by providing users an option for a virtual background) and implicitly (by excluding dark skinned people from effectively using the virtual background) incorporated into software design. Values are particularly tangible in relation to thresholds, including tolerance levels for false positives or false negatives (see also Kraemer et al., 2011).

A military example of values implemented in technology is the case of the design of the safety pin on a rifle, which embodies the value of safety. Different weapon safety mechanism designs exist, each, as it seems, having a different use or user in mind and weighing value conflicts differently, for example, a quick release mechanism over deliberate delay. The most common way if designing for values in the case of a rifle is that it has safety measures integrated with the mode selector, such as a fire selector with positions from 'safe' to 'semi-automatic' to 'full-automatic fire' (e.g. M16). Others have integrated the safety functionality in the trigger: the striker cannot move unless the safety trigger is fully depressed, which sometimes can only physically be done by a large hand. Examples of which can be found in Glock and Walther pistols. These two examples show that there are different ways of designing a value into a technology: in the first version there is an 'add-on' feature, namely the safety lever that is independent of handling of the weapon (i.e. pulling the trigger), while in the second version the safety function coincides with the handling of the weapon (i.e. pulling the trigger). Interestingly, a single design feature can both positively and negatively influence a single value for which one is seeking to account. Consider the in-built trigger safety lever which serves to prevent accidental discharges, thus prioritising safety for users and bystanders, which additionally removes the time delay associated with mode selector manipulation, which potentially deprioritises the very same value of safety for the target. The question regarding who's values to consider is as important as which values to consider.

A quick glance over the VSD literature reveals that the number of publications where practitioners report on how

[2] (1) Stakeholder Analysis, (2) Stakeholder Tokens, (3) Value Source Analysis, (4) Co-evolution of Technology and Social Structure, (5) Value Scenario, (6) Value Sketch, (7) Value-oriented Semi-structured Interview, (8) Scalable Information Dimensions, (9) Value-oriented Coding Manual, (10) Value-oriented Mockup, Prototype or Field Deployment, (11) Ethnographically Informed Inquiry regarding Values and Technology, (12) Model for Informed Consent Online, (13) Value Dams and Flows, (14) Value Sensitive Action-Reflection Model, (15) Multi-lifespan timeline, (16) Multi-lifespan co-design and (17) Envisioning Cards™ have been listed by Friedman and Hendry (2019).

they followed an actual VSD process is underrepresented and VSD literature mostly comes from the hands of academic researchers that report conceptual and/or empirical investigations, but that lack the technical step. Practical examples where values have been implemented in design have been provided above (equity in Zoom backgrounds, the safety pin in a weapon) and a recent example in electric vehicle (EV) design is where the value of an overlooked stakeholder group, namely pedestrians safety has lead governments to pass legislation mandating that EV engine sound must be substituted with artificial sound emission systems (Faas & Baumann, 2021), which could be considered a VSD approach to upgrading EVs.

Critical reflections on VSD are rather limited and mostly comprise of elaborations and updates of the VSD framework. Jacobs and Huldtgren (2018) argue against an often-identified weakness of VSD, namely that it lacks ethical commitments and objective normative power. A lack of ethical commitments can lead to a technology that is designed according to the dictates of the majority stakeholders with unintended yet severe consequences for a minority group. Jacobs and Huldtgren (2018) argue that without an explicit ethical commitment, VSD lacks a methodology for distinguishing genuine moral values from mere stakeholders-preferences and runs the risk of attending to a set of values that is unprincipled or unbounded.

Umbrello (2018) pointed out the importance of moral intuitions in determining stakeholder values in VSD, which is important for the question of where the values come from. He concludes that the VSD methodology should diminish the influence of cognitive biases with respect to moral intuitions and hence values, by adopting certain heuristic tools that are capable of doing this, thus strengthening the reliability of moral intuitions. An example of such a heuristic test is Bostrom and Ord's (2006) Double Reversal Test[3], which intends to reduce the status quo bias in judgments regarding technological innovation.

Values in autonomous systems

In this section I discuss attempts that have been made to taking into account ethical values when designing autonomous systems. At the socio-technical level, there is the Institute of Electrical and Electronics Engineers (IEEE), the world's largest professional organization for engineers, which is currently leading the way in designing professional and technical standards for engineers with the emergence of pressing questions related to autonomous systems. To this end it launched the "IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems" in 2016. The IEEE Global Initiative published a report advocating the inclusion of values in autonomous systems design, entitled "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS)" (2017b). The authors state that autonomous systems "should always be subordinate to human judgment and control. […]" (IEEE, 2017b, p. 23). The authors include the importance of cultural differences in norms and values, which is important, because values and norms are often thought of as universal in some sense (see e.g. Schwartz (2012)), while this is a contested idea in philosophy (O'Neill & Machery, 2018). Autonomous systems may be deployed over culturally and geographically dispersed areas, which makes sensitivity to such differences relevant for VSD of autonomous systems.

The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (2017b, p. 20) has articulated the following abstract ethical issues related to Artificial Intelligence/Autonomous Systems (AI/AS), which are important for our VSD for autonomous weapons overview:

1. Embody the highest ideas of human rights.
2. Prioritize the maximum benefit to humanity and the natural environment.
3. Mitigate risks and negative impacts as AI/AS evolve as socio-technical systems.

The attempt to devise ethical issues that are deemed important for the IEEE community is an important step toward realisation of value sensitive autonomous systems globally and it has inspired national initiatives, for example the Australian AI Ethics Principles (Australian Government. Department of Industry, Science, Energy and Resources, 2019). However, ideas of human rights, as predominantly ensconced in Western-developed international humanitarian law, are culturally contested, and thus the use of the term human rights might attract opposition and hence potentially hamper attempts for a global awareness for including values in the design of autonomous systems. I therefore suggest finding an internationally better appreciated term. The mitigation of risks and negative impacts is another laudable attempt for autonomous systems design proposed by the IEEE Global Initiative. However, it assumes that risks can be known, which is clearly not necessarily the case with

[3] "Double Reversal Test: Suppose it is thought that increasing a certain parameter and decreasing it would both have bad overall consequences. Consider a scenario in which a natural factor threatens to move the parameter in one direction and ask whether it would be good to counterbalance this change by an intervention to preserve the status quo. If so, consider a later time when the naturally occurring factor is about to vanish and ask whether it would be a good idea to intervene to reverse the first intervention. If not, then there is a strong prima facie case for thinking that it would be good to make the first intervention even in the absence of the natural countervailing factor." (Bostrom & Ord, 2006, p. 673).

emerging technologies. In order to further the philosophical issues related to the listed issues, the IEEE has appointed a Committee for Classical Ethics in Autonomous and Intelligent Systems in 2017. This committee explores the relevance of "established ethics systems … including secular philosophical traditions such as utilitarianism, virtue ethics, and deontological ethics and religious- and-culture-based ethical systems arising from Buddhism, Confucianism, African Ubuntu traditions, and Japanese Shinto influences … in the digital age." (IEEE, 2017a, p. 1) The committee's preliminary conclusion is that it is helpful to discuss established ethical theories when designing autonomous systems and that each society should feel free to design autonomous systems that behave in accordance with its preferred ethical theory. I agree that different societies may appreciate ethical theories and values differently, however, within societies value preferences and settling on them may change over time and in different contexts (van de Poel, 2021). The Global Initiative furthermore proposes the following approach to embedding values into Autonomous Intelligent Systems (AIS):

1. Identify the norms and values of a specific community affected by AIS.
2. Implement the norms and values of that community within AIS.
3. Evaluate the alignment and compatibility of those norms and values between the humans and AIS within that community."

The above approach is consistent with the VSD approach and the key questions pertain to identifying the "specific community affected by AIS". In view of the attempts of the IEEE society to implement norms and values of a specific community, I preliminarily discuss the proposal by Baum (2020), who discusses challenges in designing autonomous systems in a way that it acts according to the aggregate ethical views of society. In the next section I will relate this approach to an important aspect in the autonomous weapons debate, namely the Martens clause, but for now I focus on Baum's argument itself. He notes that developing AI such that it acts according to the aggregate ethical views of society underlies at least two significant lines of thinking in AI ethics. One line of thinking is "coherent extrapolated volition" (Bostrom, 2014; Yudkowsky, 2004). Coherent extrapolated volition abstains from selecting an ethical view for the initial programming and instead seeks to have the AI derive its values from the values of other ethical agents and most importantly, specifically seeks to extrapolate beyond agents' existing ethical views, essentially to figure out the views that the agents would ideally have if they were as smart as the autonomous system (Baum, 2020). The other line of thinking Baum identifies is the concept of "bottom–up" ethics (Wallach et al., 2008), through which AI is trained to deduce what is ethical through interactions with its environment and with other ethical agents. One could say that this version of bottom-up ethics tries to imitate the aggregate ethical views of society, so in a way it assumes that observable behaviour and observable responses to behaviours in society comprise ethical behaviour. In a way, the AI system derives an "ought" from the observable "is" from other ethical agents. Coherent extrapolated volition however, tries to (morally) outperform aggregate ethical views of society. It assumes that the observable "is" from other moral agents can be perfected through increased intelligence and rationality of which future AI systems are assumed to be capable. Several ethical frameworks are consistent with this approach, for example virtue ethics, as it is about an agent continuously learning and ethically improving. Contrary to bottom–up ethics (and coherent extrapolated volition) is "top–down" ethics, in which the AI is designed according to a specific ethical theory from the start and therefore does not seek to identify the views of society. Baum concludes that proposals such as coherent extrapolated volition, which in short abstains from selecting an ethical view for the initial programming and lets the AI figure out which values to derive from other ethical agents, and bottom-up ethics do not do much to resolve the important decisions to be made by the designers of ethical AI. The important ethical questions relate to decisions about whose ethics views are included, how their views are identified and how individual views are combined to a single view that will guide AI behavior. These questions need to be addressed by designers upfront and VSD is a potentially helpful framework to address the important questions identified by Baum, as they force designers to think about these questions at the start of the design process.

Another approach to including values in autonomous systems is voiced by Stuart Russel, who declared that we should impose our own values on autonomous systems and he calls this the value alignment problem (Russell, 2016; Russell et al., 2016). Accordingly, the challenge is to build autonomous systems in such a way that they maximize the realization of human values (Russell, 2016). Russel's approach to maximizing human values is a philosophically problematic statement, because it is by no means clear what such values are, as seen in the previous section, and even if we could have clarity and agreement on them, values are not fixed or static and rather responsive to context and therefore highly variable, meaning they may well change over time, so how should an autonomous system account for future human values that are currently suppressed or absent in our societies? As stated before, Russell declared the value alignment problem of building autonomous systems

with values that "are aligned with those of the human race". In fact, Russel's concern is reflected in the third bullet of the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (see previous section), which states that norms and values should be aligned in accordance with those of humanity. Peterson (2019) suggests a solution to the value alignment problem, namely to take a geometric approach. He suggests improving the methodology currently applied by computer scientists in embedding moral values in autonomous systems through representing moral principles as conceptual spaces, i.e. as Voronoi tessellations[4] of morally similar choice situations located in a multidimensional geometric space (2019). I acknowledge that in reality it is very difficult or perhaps impossible to translate values into code, I appreciate that within the limited context of a few pre-given principles, Peterson's solution provides the designers of autonomous systems a workable methodology relatively easy to implement and translate into a machine-interpretable model that provides a heuristic on how to harmonize values.

I discussed several attempts to including values into autonomous systems design. My final attention goes to the socio-technical dimension. Initiatives such as IEEE (see above), including its recent P7000 Standards series (2021) may inform the organizational structure of organizations that develop autonomous systems, so the attention to values is captured in manuals, handbooks and work routines. However, equally important, is that values are part of work floor culture and discussing and questioning values in the face of changing stakeholders or user contexts, is encouraged and reflected in reward systems and appropriate reporting lines that allow for different facets of the value sensitive design process.

Values in autonomous weapons systems

In the previous section I discussed several suggestions on including values in autonomous systems in general. In this section, I critically discuss literature on autonomous weapon systems in light of VSD.

Autonomous weapons systems can be autonomous in multiple ways. One distinction that can be drawn is between an autonomous 'input' side, which relates to data gathering and target selection, while the other side is the 'output' side, namely an autonomous process related to deciding when and how to attack a target. I understand the compilation of data to be a value laden process, since it includes assigning weight and priority to the data sources. Values are thus embedded in the process of gathering data, as much as in the pre-programmed actions, such as lethal attacks, that the autonomous system performs based on the data.

The values in pre-programmed actions are relevant to Baum's (2020) view on how to include ethics in autonomous systems, which I briefly addressed in the previous section. Baum argues for designing an autonomous system in a way that it acts according to the aggregate ethical views of society. Baum's view is particularly relevant to the Martens clause, which is a provision in international humanitarian law stating that "if there is no specific law on a topic, civilians are still protected by the principles of humanity and dictates of public conscience".[5] Some authors argue that fully autonomous weapons contravene the Martens Clause by pointing to the publicly voiced opposition to fully autonomous weapons by faith leaders, scientists, tech workers, civil society organizations (e.g. Lin (2015); Docherty (2020)). Baum's approach to developing AI in a way that encourages actions according to the aggregate ethical views of society could be interpreted as a consensus between those that point out the importance of the Martens clause (i.e. listen to the public if the law does not suffice) and those that argue for the military and strategic need for autonomous weapons. Bottom–up ethics would be consistent with the Martens Clause, because it is based on learning from other ethical agents and will eventually be some aggregation representing a particular (part of) society. Baum's conclusion is particularly relevant for VSD of autonomous weapons systems: "They are inherently decisions that must be made by AI designers - one cannot "let the AI figure it out", because the decisions concern *how* (italics mine) AI would figure it out. Focus should likewise be on the important decisions, not on whether AI uses some sort of social choice ethics." (Baum, 2020, p. 175). I concur with Baum in this respect, meaning that ethicists should work with designers of autonomous weapons systems on preliminary normative questions regarding the reasons as to *why* one would want autonomous weapons systems in the first place and secondly *how* to design them in a more ethical way. One of the difficulties is to bring ethicists and designers together to discuss, in a critical yet helpful manner, what can and cannot be done to 'make autonomous weapons behave (more) ethical'. VSD can serve this purpose, so long as designers and ethicists are open about the normative assumptions and ethical commitments they adhere to. This does not mean that there is a

[4] "The concept of a Voronoi region is a simple but intuitively appealing one. Given a finite set of distinct, isolated points in a continuous space, we associate all locations in that space with the closest member of the point set. The result is a partitioning of the space into a set of Voronoi regions. The mathematical theory of Voronoi regions has successfully been applied in many areas, including philosophy (notably by Peter Gärdenfors in his works on "conceptual spaces")." (Lokhorst, 2018).

[5] Convention (II) with Respect to the Laws and Customs of War on Land and its Annex: Regulations concerning the Laws and Customs of War on Land, The Hague, adopted July 29,1899, entered into force September 4, 1900, pmbl., para. 8; Protocol I, art. 1(2).

need to agree on these assumptions and commitments, but it helps to identify common concerns and a shared understanding of the environment in which autonomous weapons are likely to be deployed to understand what is exactly at stake. Another way to put this, is in the words of Jacobs and Hultgren (2018): "In a context where people with various disciplinary backgrounds, interests, and priorities have to work together, which often is the case in design-contexts, convergence on the practical level is crucial to come to joint decisions."

I will now discuss several attempts found in literature that bring together defence, industry and societal stakeholders to discuss values in autonomous weapons.

I start with the US Defence Science Board Task Force Report (2012) on "The Role of Autonomy in DoD Systems", which states that it is important for commanders to

> "maintain the human-machine collaboration needed to execute their mission, which is frequently handicapped by poor design. A key challenge facing unmanned system developers is the move from a hardware-oriented, vehicle centric development and acquisition process to one that addresses the primacy of software in creating autonomy. For commanders and operators in particular, these challenges can collectively be characterized as a lack of trust that the autonomous functions of a given system will operate as intended in all situations." (DoD DSC, 2012, p. 2).

Thus, the value of trust is recognized by defence personnel as of utmost importance in autonomous weapons systems. The importance of the value of trust in an operational setting is also pointed out by Ekelhof (2019). I acknowledge with Ekelhof the need to thoroughly understand the context of application of autonomous weapons and that tacit rules and legal boundaries are inherently present, as I have made clear at the start of this paper, and is addressed in more detail by Umbrello (2021). Autonomous weapons systems on the battlefield are or will not be merely systems that make a kill/no kill decision. As in current military operations, many decisions have been made before one arrives at the kill/no kill decision and it is likely that this will hold for autonomous weapons systems too. Ekelhof makes a helpful analysis of a 'traditional' lethal targeting routine of F16 fighter pilots to explain that a targeting process is a highly distributed process in which several steps exist and several experts are involved before the weapons are released and a target is engaged. She states that "even though the operator did not partake in collecting and analyzing […] information, there is a matter of trust that the targets that they are tasked with comply with the law and rules of engagement and are in line with the mandate" (Ekelhof, 2019, p. 346). In this one quote, there is a direct link to the legal aspect and the trust aspect in the context targeting: the operator *trusts* that others have delegated a *lawful* task to them. "The information about the lawfulness of the action thus largely depends on the operator's trust in his or her superiors in the chain of command (to provide proper briefing materials and conduct target validation), the F16board computer (suggesting the appropriate time for weapons' release) and the weapon's guidance system (navigating the munitions to the target)." (Ekelhof, 2019, p. 346). These are important observations for a VSD perspective on autonomous weapons, because one needs to account for the values of trust and lawfulness in the technological design of the system and moreover, the quote makes clear that a narrow focus on the operator controlling the weapon during mission execution does not pay due attention to the embeddedness of the technology in a specific structure, where technology and humans play their respective roles, as also Elish (2017) has pointed out. The operator, in whatever capacity or remoteness, acts in a specific normative structure (see De Vries and Jochemsen 2019; van Burken and Bezooijen 2015)) and therefore the values that are built in the autonomous systems should account for these normative structures, or else there is a risk that the technology will be rejected altogether.

Verdiesen (2017) first suggested VSD for the context of autonomous weapons systems and Santoni de Sio and Van den Hoven (2012) mention VSD in the context of meaningful human control over autonomous systems. Verdiesen, Santoni de Sio and Dignum (2018) studied the *perception* of moral values in the deployment of autonomous weapons systems. The value theories of Schwartz, Friedman and Kahn, Beauchamp and Childress were mapped against the autonomous weapons debate and surveys were held amongst Dutch military personnel and civilians. They found the following values to be important in the autonomous weapons debate: blame, trust, harm, human dignity, confidence, expectations, support, fairness and anxiety. A careful reading of these findings shows that not all the words listed are actual values, as some are considered factors relating to values. The research by Verdiesen and colleagues points us in a fruitful direction for future research on VSD and could serve as a precursor for doing VSD with autonomous weapons developers. Verdiesen and colleagues have certainly contributed to a greater understanding of values in relation to autonomous weapons, however, one weakness of their approach relates to the use of a sample scenario from a warzone in order to derive perceptions of values. Even though the authors are alert with regard to the neutral description of the agency question in the vignette, the situation they describe is not neutral with respect to its different respondents. In the vignette they describe a fast car approaching behind a mountain range and this car potentially poses a

threat, because it is identified as carrying weapons and the driver is recognized as an insurgent. The problem with this vignette is that a 'fast car behind a mountain range' is not immediately appealing or recognized for most Dutch citizens of which the researchers took a sample, but it is a recognizable situation for military personnel who have served in an international mission. Therefore, the vignette with the car approaching behind a small mountain range may immediately recall feelings of threat and danger for military personnel that have served in a mission, whereas the average Dutch person may think of a geographically, emotionally and morally distant scenario. Any research on values and autonomous weapons that distils values through scenario descriptions must account for the potential value-ladenness of the scenario description itself. It may contain important contributors to whether someone trusts the decision of an autonomous weapon or not, and related to this, evoke different value perceptions. In short, a VSD approach that deploys scenarios for mining values from stakeholders must account for sociological and psychological factors related to the use of force (which may greatly differ between countries where citizens highly trust their governments in the use of force or where there is no or low trust in military force), see also Roeser (2010) for a discussion on technology, ethics and emotions.[6]

Verdiesen and Dignum (2022) presented further empirical research on values related to autonomous weapons. They use value deliberation techniques on two participant groups, presenting them with scenarios that contain realistic levels of detail and information about the facets of the autonomous systems. They did not test their scenarios on lay people, meaning the values are coloured by professional experiences, such as military experience, or expert knowledge about autonomous systems, however they gathered highly valuable data and observations regarding values in autonomous weapons design.

Umbrello (2019) and Umbrello et al. (2020) assess and engage the current arguments for and against the use of lethal autonomous weapons and they suggest a VSD approach for developing lethal autonomous weapons that make decisions within the bounds of their ethics-based code, in order to potentially raise social acceptance of such weapons. Interestingly, Umbrello (2019) suggests to include *mainly* military stakeholders' values in the design of autonomous weapons, which weakens their claim that VSD potentially increases social acceptance, as one would expect that including societal values, rather than military values, would potentially raise the social acceptance of autonomous weapons. The authors claim that 'ethical' should be used in a pragmatic way, "such that an ethical LAW [Lethal Autonomous Weapon] is one that functions in accordance with the LoW [Laws of War] and RoE [Rules of Engagement]" (Umbrello et al., 2020, p. 276). I agree that laws of war and rules of engagement reflect certain moral values and an attempt to program these into autonomous weapons is a first step in VSD of autonomous weapons systems. However, confining ethics to adherence to laws of war and the rules of armed conflict that can be programmed into autonomous weapons systems, suggests that autonomous weapons themselves are ethically neutral and only when we program into the system the laws of war or the rules of engagement, the system becomes an ethical system. In this reading, their remark is inconsistent with the non-neutrality thesis to which the authors claim to adhere to, by stating that "the essentialist conception of technology as a value-neutral tool has long been shown to be a misguided one". (Umbrello et al., 2020, p. 278) In other words, Umbrello and colleagues claim that the non-neutrality thesis is a misguided one, but at the same time they seem to suggest that an autonomous weapon is ethically neutral so long as no ethical and legal laws are programmed into the system. This suggests a technological fix, rather than an acknowledgement that the autonomous system is not value neutral with respect to morally important values, irrespective of the legal and ethical rules that are being programmed into the autonomous system.

Nevertheless, Umbrello and colleagues point out the importance of the LoW and ROEs in developing autonomous weapons in a value sensitive manner, which is a highly important minimum requirement for VSD of autonomous weapons. A minor critique on their attempt to conceptualize autonomous weapons, is that they seem to fall prey to an overly simplistic understanding of how LoW and ROEs can be programmed into an autonomous system. The reality of targeting and military operations is highly complex, as Ekelhof (2019) has shown. I agree with Umbrello and colleagues that considering existing legal frameworks around the use of force is a good starting point for designing autonomous weapons. However, assuming that autonomous weapons can be programmed such that they replicate international humanitarian law and ROEs neglects the complexity of technical translation, formalization and operationalisation of the rules into autonomous weapons. Formalizing legal rules into software algorithms is a technological design step, which requires sensitivity to the reality of military operations and it involves decisions about the sensitivity, offsets and thresholds of sensors that determine the reliability of the data that serves as input for selecting and engaging targets by an autonomous weapon. The importance of the correctness of decisions needs much more attention than it has thus

[6] According to Roeser (2010), technologies can trigger emotions, including fear and indignation, which often leads to conflicts between stakeholders. She argues that moral emotions can play an important role in judging ethical aspects of technological risks, such as justice, fairness, and autonomy.

far had in the current debates on VSD of autonomous weapons. I argue that merely programming legal or ethical rules into autonomous systems are not the only things that make the system value-laden. Instead, the value ladenness is also embedded in the number and nature of sensors (e.g. prioritizing visual data over voice), programming of thresholds (e.g. prioritizing false positives over false negatives), sensitivities, weights, battery life, etcetera, which are ethically relevant design choices.

What are the potential threats and opportunities specific to VSD for autonomous weapons?

Autonomous weapons design is based on the kind of technology that attracts a rich public debate, which is a potential opportunity for VSD for autonomous weapons. Given the current interest of politicians, pressure from the international community and continued coverage by media it is vitally important for the developers of autonomous weapons to take seriously the concerns of different representatives in society. As has become clear in recent years, failure to listen to societal (ethical) concerns regarding a technology may lead to a rejection of the technology altogether.[7] Taebi et al. (2014) suggest that relevant public values can be extracted from these public debates, which in turn can inform the design and policy regarding controversial technologies. Autonomous weapons are one of the most controversial technologies in recent history and a VSD of autonomous weapons should therefore be informed by the actual public debates, for which there are ample opportunities. Zolyomi (2018) and Briggs and Thomas (2010) suggests that tapping into social media during VSD research is a fruitful manner to get public values on the table. Another resource for extracting values are the AI ethics guidelines and principles that have emerged from e.g. NATO (2021), United States of America Defence Innovation Board (Defense Innovation Board, 2019), United Kingdom Department of Defence (2022), Australian Defence Science and Technology Group (Kate Devitt et al., 2021).

One of the challenges that comes with a complex design process such as VSD of autonomous weapons systems is the socio-technical dimension, such as how to harmonize the timelines and agendas of the different parties in the design process. This may not always be possible for practical or budgetary reasons and it is likely that the party who does the conceptual VSD work may be reassigned to a different task by the time there is a need for a second or third round of conceptual or empirical work after the technical implementation is finalized, which blocks the iterative nature that is assumed by VSD. Related to this is that Defence acquisition processes, which provide great opportunities to encourage respect for certain values, are often times lengthy and do not work well for an iterative process.

The second challenge is related to distilling values. Different scenarios for mining values are possible and in the case of VSD of autonomous weapons, ideally, stakeholders from inside and outside industry and defence should partake. The challenge is to find a fruitful mode to discuss autonomous weapons systems and the relevant values, but this may be problematic because of the classified status of such technologies, or because of the competitiveness of designing autonomous systems.

Another challenge when doing empirical research on values in autonomous systems, is the danger of considering the issue in a biased manner, for example when the questions regarding values in autonomous weapons reach an audience that may have never been in an armed conflict, or that either live in a high trust or low trust society (where governments are trusted easily or not at all), as discussed in the previous section.

Yet another challenge is that the debates surrounding the research on autonomous weapons systems can become polarized very quickly. Controversial technologies tend to attract a strong voice from opponents, a less strong voice from proponents, and almost a silent majority. In such a polarized debate on autonomous weapons, it is difficult to get stakeholders from the military and some pressure groups from the societal domain around the table, potentially for fear that they will be unfavourably labelled or treated.

A further challenge for pursuing the VSD of autonomous weapons is the classified nature and confidentiality regarding the technologies under development. It may become difficult to do the iterative empirical – conceptual - technical work, because at the technical level security restrictions will often be so tight as to preclude meaningful engagement. VSD discussions with ideally many stakeholders involved may be done on a very abstract level, because for reasons of confidentiality it is not possible to discuss concrete technologies. A way to mitigate this issue is to discuss hypothetical cases, or to break up discussions with different stakeholders, so as to ensure there is no cross-over of classified information between stakeholder groups.

A final challenge is that there is no clear understanding of what is meant by “autonomous system”, or “autonomous weapon”, so this may hamper the discussions (see also my observation in Sect. 2). This issue is not easily resolved, it may even be undesirable to have an agreement on the definition of an autonomous weapons system, because there will always be edge cases that fall between the cracks, and settling on a definition means that currently existing weapons

[7] https://www.complianceweek.com/opinion/top-ethics-and-compliance-failures-of-2018/24720.article; Tannenwald, N. (2005). Stigmatizing the Bomb: Origins of the Nuclear Taboo. International Security, 29(4), 5–49. http://www.jstor.org/stable/4137496.

systems may need to undergo new review processes as they now fall under the definition of 'autonomous weapons system' (Horowitz, 2016; Taddeo & Blanchard, 2022). Nevertheless, one can focus discussions by focus on the characteristics and attributes of the systems under consideration rather than value-laden definitions.

Conclusion and recommendation

I discussed a VSD approach for addressing the question of values in autonomous systems and in particular in autonomous weapons systems. I discussed different approaches to embedding values in autonomous systems, as they were found in literature. Some of these approaches seem fruitful for responsible technology development, such that moral values are taken into account. For example, Peterson's geometrical approach to the value alignment problem (2019) and Verdiesen et al. (2019) who did a survey on value perception regarding autonomous weapons. Trust appeared to be an important value in literature on values in autonomous systems.

I found limited literature where authors reported on how they followed an actual VSD process, that was called VSD and included conceptual, empirical and technical steps, resulting in a certain (socio-)technological design in general. I found none on autonomous systems specifically, probably because researchers do not often engage in an actual VSD process on the one hand, and practitioners that follow the VSD methodology may not report on it in academic literature on the other hand.[8] There is a lack of account of the usefulness and practical feasibility of VSD for the actual design of autonomous systems. To better inform the ideas developed here, I intend to apply VSD to an autonomous weapons design process and report on the values that have been considered and implemented in the design of actual technological prototypes or design phases, including a discussion on the actual process of VSD of autonomous weapons. I recognize that this may be a challenge for several reasons, which I discussed in the previous section.

I suggested the importance of the socio-technical dimension of value sensitive design, captured in e.g.IEEE initiatives. They may inform systems engineering practices and in this way they carry and embed important values that may be reflected in the artifacts that are being designed in these practices.

I furthermore conclude that VSD of autonomous weapons can be used to gain societal support for development of autonomous weapons, but this requires a careful and inclusive selection of stakeholders and this can be, for example, through Critical Systems Heuristics (Ulrich & Reynolds, 2010). The next step is to (1) outline a VSD methodology that includes aforementioned critical stakeholder selection, (2a) hold actual stakeholder meetings and (2b) do content analysis from public sources and social media to elicit values, and (3) engage in a technological design process with actual developers of autonomous weapons. In order to enhance the chance of success for the next steps, and in particular the technical design step, it is important to see how VSD can complement or support existing systems design methodologies that system designers may be familiar with, such as spiral development, DevSecOps and agile development.

Acknowledgements The author(s) like to thank two anonymous reviewers for their helpful comments.

Funding This research received funding from the Australian Government through Trusted Autonomous Systems, a Defence Cooperative Research Centre funded through the Next Generation Technologies Fund. The views and opinions expressed here are those of the author, and do not necessarily reflect the views of the Australian Government or any other institution. Open Access funding enabled and organized by CAUL and its Member Institutions.

Declaration

Conflict of interest The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Advisory council on International Affairs (2015). *Autonomous Weapon Systems: The Need for Meaningful Human Control* (No. 97 AIV / No. 26 CAVV, October 2015).

Arkin, R. C. (2010). The case for ethical autonomy in Unmanned Systems. *Journal of Military Ethics*, *9*(4), 332–341. https://doi.org/10.1080/15027570.2010.536402

Assuring Body of Knowledge. (n.d.). *Assuring Body of Knowledge Definitions*. Assuring Autonomy International Programme. Retrieved July 7 (2020). from https://www.york.ac.uk/assuring-autonomy/body-of-knowledge/definitions/

[8] I recognize that in this article I have not engaged in a VSD process of autonomous weapons systems either, hence the addition to the title: primer".

Australian Government. Department of Industry, Science, Energy and Resources (2019). *AI Ethical Principles*. https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles

Baum, S. D. (2020). Social choice ethics in artificial intelligence. *AI & SOCIETY*, *35*(1), 165–176. https://doi.org/10.1007/s00146-017-0760-1

Boshuijzen-van Burken, C. (2016). Beyond technological mediation: a normative practice approach. *Techné*, *20*(3), 177–197. https://doi.org/10.5840/techne201671949

Boshuijzen-van Burken, C., & Bezooijen, B. (2015). Morally Responsible Decision Making in Networked Military Operations. In B.-J. Koops, I. Oosterlaken, H. Romijn, T. Swierstra, & J. van den Hoven (Eds.), *Responsible Innovation 2: Concepts, Approaches, and Applications* (pp. 265–282). Springer International Publishing. https://doi.org/10.1007/978-3-319-17308-5_14

Boshuijzen-van Burken, C. (2021). Modern Military Operations: A Normative Practice Approach to Moral Decision Making. In I. Management Association (Ed.), *Research Anthology on Military and Defense Applications, Utilization, Education, and Ethics* (pp. 522–535). IGI Global. https://doi.org/10.4018/978-1-7998-9029-4.ch028

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies* (First edition). Oxford University Press.

Bostrom, N., & Ord, T. (2006). The reversal test: eliminating status quo bias in applied ethics. *Ethics*, *116*(4), 656–679. https://doi.org/10.1086/505233

Boyd, K. (2022). Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2069–2082. https://doi.org/10.1145/3531146.3534626

Briggs, P., & Thomas, L. (2015). An Inclusive, Value Sensitive Design Perspective on Future Identity Technologies. *ACM Transactions on Computer-Human Interaction*, *22*(5), 23:1–2328. https://doi.org/10.1145/2778972

Costley, D. (2020, October 27). Zoom's Virtual Background Feature Isn't Built for Black Faces. https://onezero.medium.com/zooms-virtual-background-feature-isn-t-built-for-black-faces-e0a97b591955

De Vries, M. J., & Jochemsen, H. (Eds.). (2019). *The Normative Nature of Social Practices and Ethics in Professional Environments*. IGI Global. https://doi.org/10.4018/978-1-5225-8006-5

Defense Innovation Board (2019). AI principles: Recommendations on the ethical use of Artificial Intelligence by the Department of Defense. *Defense Innovation Board*. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

Docherty, B. (2020). The Need for and Elements of a New Treaty on Fully Autonomous Weapons. *Proceedings of Rio Seminar on Autonomous Weapons Systems, 20 February 2020*. Rio Seminar on Autonomous Weapons Systems, Rio de Janeiro. https://www.hrw.org/sites/default/files/media_2020/06/202006arms_rio_autonomous_weapons_systems_2.pdf

DoD DSC (2012). *The Role of Autonomy in DoD Systems*. Department of Defence Defense Science Board. https://fas.org/irp/agency/dod/dsb/autonomy.pdf

Dooyeweerd, H. (1953). *A new critique of theoretical thought: vol. I–V*. The Presbyterian and Reformed Publishing Company.

Ekelhof, M. (2019). Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation. *Global Policy*, *10*(3), 343–348. https://doi.org/10.1111/1758-5899.12665

Elish, M. C. (2017). Remote split: a history of US drone operations and the distributed labor of war. *Science Technology & Human Values*, *42*(6), 1100–1131.

Faas, S. M., & Baumann, M. (2021). Pedestrian assessment: is displaying automated driving mode in self-driving vehicles as relevant as emitting an engine sound in electric vehicles? *Applied Ergonomics*, *94*, 103425. https://doi.org/10.1016/j.apergo.2021.103425

Floridi, L., & Sanders, J. W. (2004). On the morality of Artificial Agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Friedman, B. (1996). Value-sensitive design. *ACM Interactions*, *3*(6), 17–23.

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: shaping technology with moral imagination*. Mit Press.

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

Friedman, B., & Kahn, P. (2003). Human Values, ehics and design. In *The human-computer interaction handbook* (pp. 1177–1201). https://brandorn.com/img/writing/tech-ethics/human-values-ethics-and-design.pdf

Friedman, B., Kahn, P., & Borning, A. (2002). Value sensitive design: Theory and methods. *University of Washington Technical Report*, 02–12.

Friedman, B., Kahn, P., & Borning, A. (2006). Value sensitive design and information systems. In P. Zhang, & D. Galletta (Eds.), *Human-Computer Interaction in Management Information Systems: foundations* (pp. 348–372). M.E. Sharpe.

GGE LAW (2019). *Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems* (CCW/GGE.1/2019/3). https://undocs.org/en/CCW/GGE.1/2019/3

Horowitz, M. C. (2016). Why words Matter: the Real World Consequences of defining Autonomous Weapons Systmes. *Temple International & Comparative Law Journal*, *30*, 85.

ICRC (2019). *Artificial intelligence and machine learning in armed conflict: A human-centred approach*. https://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach

IEEE (2017a). *Classical Ethics in A/IS*. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_classical_ethics_ais_v2.pdf

IEEE (2017b). *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2*. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

IEEE. (2021). IEEE 7000 – 2021—IEEE standard model process for addressing ethical concerns during System Design. *(IEEE Standards ISBN*, *9781504476874*, 9781504476881, 9781504479356. IEEE Computer Society.https://standards.ieee.org/standard/7000-2021.html

Jacobs, N., & Huldtgren, A. (2018). Why value sensitive design needs ethical commitments. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-018-9467-3

Jenkins, K. E. H., Spruit, S., Milchram, C., Höffken, J., & Taebi, B. (2020). Synthesizing value sensitive design, responsible research and innovation, and energy justice: a conceptual review. *Energy Research & Social Science*, *69*, 101727. https://doi.org/10.1016/j.erss.2020.101727

Kate Devitt, M., Gan, J., Scholz (2021). *A method for ethical AI in Defence* (DSTG-TR-3786). Australian Government Department of Defence. https://apo.org.au/node/311150

Kraemer, F., van Overveld, K., & Peterson, M. (2011). Is there an ethics of algorithms? *Ethics and Information Technology*, *13*(3), 251–260. https://doi.org/10.1007/s10676-010-9233-7

Lin, P. (2015). *The right to life and the Martens Clause*. Convention on Certain Conventional Weapons (CCW) meeting of experts on lethal autonomous weapons systems (LAWS), at United Nations in Geneva, Switzerland on 13–17 April 2015. http://cyberlaw.stanford.edu/files/publication/files/ccw_testimony.pdf

Lokhorst, G. J. C. (2018). Martin Peterson: The Ethics of Technology: A Geometric Analysis of Five Moral Principles: Oxford University Press, 2017, 252 pp, USD 74.00 (hbk), ISBN: 9780190652265. *Science and Engineering Ethics*, *24*(5), 1641–1643. https://doi.org/10.1007/s11948-017-0014-0

Miller, B. (2020). Is Technology Value-Neutral? *Science, Technology, & Human Values*, 016224391990096. https://doi.org/10.1177/0162243919900965

NATO. *NATO Review—An Artificial Intelligence Strategy for NATO*. NATO Review (2021, October 25). https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html

Nickel, P. J. (2015). Design for the Value of Trust. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of Ethics, Values, and Technological Design* (pp. 551–567). Springer Netherlands. https://doi.org/10.1007/978-94-007-6970-0_21

O'Neill, E., & Machery, E. (2018). The Normative Sense. In A. Zimmerman, K. Jones, & M. Timmons (Eds.), *The Routledge Handbook of Moral Epistemology* (1st ed., pp. 38–56). Routledge. https://doi.org/10.4324/9781315719696-3

Peterson, M. (2019). The value alignment problem: a geometric approach. *Ethics and Information Technology*, *21*(1), 19–28.

Roeser, S. (Ed.). (2010). *Emotions and risky technologies*. Springer.

Russell, S. (2016). Should we fear Supersmart Robots? *Scientific American*, *314*(6), 58–59. https://doi.org/10.1038/scientificamerican0616-58

Russell, S., Dewey, D., & Tegmark, M. (2016). Research Priorities for Robust and Beneficial Artificial Intelligence. *ArXiv:1602.03506 [Cs, Stat]*. http://arxiv.org/abs/1602.03506

de Santoni, F., & van den Hoven, J. (2018). Meaningful human control over Autonomous Systems: a philosophical account. *Frontiers in Robotics and AI*, *5*, 15. https://doi.org/10.3389/frobt.2018.00015

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, *2*(1), 11. https://doi.org/10.9707/2307-0919.1116

Strawser, B. J. (2010). Moral predators: the duty to employ uninhabited aerial vehicles. *Journal of Military Ethics*, *9*(4), 342–368.

Taddeo, M., & Blanchard, A. (2022). A comparative analysis of the definitions of Autonomous Weapons Systems. *Science and Engineering Ethics*, *28*(5), 37. https://doi.org/10.1007/s11948-022-00392-3

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

UK Ministry of Defence. (2022). *Ambitious, safe, responsible. Our approach to the delivery of AI-enabled capability in Defence*. UK Ministry of Defence.

Ulrich, W., & Reynolds, M. (2010). Critical systems heuristics. *Systems approaches to managing change: a practical guide* (pp. 243–292). Springer.

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.

Umbrello, S. (2019). Lethal Autonomous Weapons: Designing War Machines with values. *Delphi: Interdisciplinary Review of Emerging Technologies*, *1*(2), 30–34.

Umbrello, S. (2021). Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: a two-tiered approach. *Ethics and Information Technology*, *23*(3), 455–464. https://doi.org/10.1007/s10676-021-09588-w

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

Umbrello, S., & van de Poel, I. (2020). Mapping Value Sensitive Design onto AI for Social Good Principles. *Preprint*.

van de Kaa, G., Rezaei, J., Taebi, B., van de Poel, I., & Kizhakenath, A. (2020). How to weigh values in Value Sensitive Design: a best worst Method Approach for the case of Smart Metering. *Science and Engineering Ethics*, *26*(1), 475–494. https://doi.org/10.1007/s11948-019-00105-3

van de Poel, I. (2021). Design for value change. *Ethics and Information Technology*, *23*(1), 27–31. https://doi.org/10.1007/s10676-018-9461-9

van de Poel, I., & Royakkers, L. M. M. (2011). *Ethics, technology, and engineering: an introduction (paperback)* (67 vol.). Wiley -Blackwell.

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., Vermaas, P. E., & van de Poel, I. (2015). *Handbook of ethics, values, and technological design*. Springer Netherlands: Imprint: Springer.

Van Wynsberghe, A. (2013). Designing robots for care: care centered value-sensitive design. *Science and Engineering Ethics*, *19*(2), 407–433.

Verdiesen, I. (2017). How do we ensure that we remain in control of our autonomous weapons? *AI Matters*, *3*(3), 47–55. https://doi.org/10.1145/3137574.3137585

Verdiesen, I., & Dignum, V. (2022). Value elicitation on a scenario of autonomous weapon system deployment: a qualitative study based on the value deliberation process. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00211-2

Verdiesen, I., de Sio, F. S., & Dignum, V. (2019). Moral values related to Autonomous Weapon Systems: an empirical survey that reveals Common Ground for the ethical debate. *IEEE Technology and Society Magazine*, *38*, 34–44.

Vermaas, P. E., Hekkert, P., Manders-Huits, N., & Tromp, N. (2015). 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. 179–201). Springer Netherlands. https://doi.org/10.1007/978-94-007-6970-0_10

Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *AI & SOCIETY*, *22*(4), 565–582. https://doi.org/10.1007/s00146-007-0099-0

Winkler, T., & Spiekermann, S. (2021). Twenty years of value sensitive design: a review of methodological practices in VSD projects. *Ethics and Information Technology*, *23*, 17–21. https://doi.org/10.1007/s10676-018-9476-2

Wolterstorff, N. (1983). *Until justice and peace embrace: the Kuyper lectures for 1981 delivered at the Free University of Amsterdam*. Eerdmans Pub Co.

Yudkowsky, E. (2004). *Coherent extrapolated volition*. Singularity Institute for Artificial Intelligence. https://intelligence.org/files/CEV.pdf

Zolyomi, A. (2018). Where the stakeholders are: tapping into social media during value-sensitive design research. *Ethics and Information Technology*, 1–4. https://doi.org/10.1007/s10676-018-9475-3