PhD DISSERTATIONS OF THE
NORTHWESTERN ITALIAN PHILOSOPHY CONSORTIUM – FINO

Towards a Value Sensitive Design Framework for Attaining Meaningful Human Control over Autonomous Weapons Systems

Steven Umbrello (XXXIV Cycle)

2021

Examination Committee:

Prof. Maurizio Balistreri	[thesis supervisor] University of Turin
Prof. Stefan Lorenz Sorgner	John Cabot University
Prof. Alberto Eugenio Ermenegildo Pirni	Sant'Anna School of Advanced Studies
Prof. dr. Ir. Ibo van de Poel	Delft University of Technology
Prof. Marcello Chiaberge	Polytechnic University of Turin

Northwest Philosophy Consortium FINO

TOWARDS A VALUE SENSITIVE DESIGN FRAMEWORK FOR ATTAINING MEANINGFUL HUMAN CONTROL OVER AUTONOMOUS WEAPONS SYSTEMS

DISSERTATION

To obtain
the degree of doctor at the Consortium FINO
XXXIV Cycle
2018-2021

By

Steven Umbrello

This dissertation has been approved by:

Coordinator:

Prof. Anna Elisabetta Galeotti	University of Eastern Piedmont

Supervisor:

Prof. Maurizio Balistreri	University of Turin

Committee Members:

Prof. Stefan Lorenz Sorgner	John Cabot University
Prof. Alberto Eugenio Ermenegildo Pirni	Sant'Anna School of Advanced Studies
Prof. dr. Ir. Ibo van de Poel	Delft University of Technology
Prof. Marcello Chiaberge	Polytechnic University of Turin

Names: Umbrello, Steven, 1993-author.
Title: Towards a Value Sensitive Design Framework for Attaining Meaningful Human Control over Autonomous Weapons Systems / Steven Umbrello
Description: Torino, TO : Consorzio FINO, [2021] | Includes bibliographical references

Acknowledgements

Coming to the end of my PhD journey is a strange thing. I thankfully did not experience any of the horrors of isolation, mental strain, or issues with supervision or completion that I have read about or witnessed with my colleagues. In fact, the experience of completing my PhD when and where I have chosen to undertake it has been the most important decision I have ever made.

Although it is not uncommon for students to move away from their birthplace to undertake their studies, my PhD was the first time I ever lived away from my nation of birth. It was here, in Turin, that I became a homeowner for the first time, learned the art of homemaking, and fell in love. The amazing people and experiences that I have been fortunate enough to meet are responsible for where I am now and each of those seemingly tiny interactions culminated in what follows in these pages in one way or another.

Philosophy is strange subject to devote one's life to. In one way or another, we all do philosophy. The way we wake up in the morning, how we speak to strangers and to loved ones, and how we make difficult decisions in seemingly unfair circumstances. All of these experiences help us to become whoever we are in any given moment. I have done my best to give this thesis a human touch, to make it of value to those who will take the time to read it, and perhaps have some greater purpose in the world of things to come.

All in all, this is a project of collaboration, for no single person is truly self-made, they surge into the world with the help of those they find themselves surrounded by, who give them small, yet profound words of comfort and advice and support them even with they seem alone.

Maurizio Balistreri, thank you for accepting me as your first PhD student. When I met you I didn't even know our research backgrounds would be so similar, nor did I know our personal characters would be so in sync. You never pressured me unduly, nor expected me to conform to any burdensome norms. Thank you for giving me the freedom to purse the line of inquiry that I found most interesting and the following pages are shaped most fundamentally by your constant, open, and honest guidance over the course of the last three years. Thank you for teaching me to stand my ground, to not give into the politics of the academy, and to pursue the philosophy that is important *per se*.

I would like to thank the Global Catastrophic Risk Institute, particularly Seth Baum. In 2013 when I reached out to become affiliated with his organisation he welcomed me with open arms. It was under Seth's mentorship that I became familiar with the world of real academic research, something lacking in the academy during that point in my studies. Thank you for seeing my potential as a burgeoning researcher, guiding me towards the best practices in scholarship, and for providing me with invaluable experiences in publishing that have only helped my career path.

Thank you to the Institute for Ethics and Emerging Technologies, particularly James Hughes and Marcelo Rinesi who have given me nothing but support in my research and providing me with a platform that has done nothing less than open doors for me along the way. James, thank you in particular for always being available as an open ear to give me guidance as to my career path, I owe my current state to you.
To Ibo van de Poel and Jeroen van den Hoven I owe my thanks, for they were the ones, albeit unknowingly, who inspired me to pursue the ethics of technology and whose works were where I first encountered the Value Sensitive Design approach that has since been the cornerstone of my research. Their names have and continue to be synonymous with celebrity in my field of study, and as unthinkable as it was to believe I would shortly after work closely with them, alas I have had such a privilege. Without you, this opus would be nonexistent.

To Louise Chapman I owe thanks for her skill, expediency, and attentiveness in making thorough copy-edits to this thesis. To those who read this and happen to be impressed with its flow and readability, the laurels are hers.

Finally, I would like to thank my family. Mom and dad, I know that my leaving has and continues to be hard on you and that you have missed me very much in my absence. I know that you were hesitant to let me go, and did you best to make me stay. Despite all, you have been supportive throughout this journey, not only financially but emotionally. Without either of these the following pages would be nonexistent. You did your best with your means and at great distance to help me to create a safe and comfortable environment to allow this work to emerge naturally, for me to explore new boundaries, and surge into my career. I would like to thank my grandparents who raised me with their values, culture, and language, of which have been invaluable in my transition to a new life in a new country. I would like to thank in particularly my paternal grandfather for giving me particular council that has been my guiding star every day since I left: that all I need is to wake up each morning with a heart full of love and affection, and that I would be fine if I did so. You were right.

Note on the Cover Art

The title of this essay reflects the focal points of the author's research and design process.
The cover artwork is focused on the concept of "frameworks": a series of circle overlaps, starting from the same pivotal point, creating a framework and a multitude of shapes which are all connected to each other.

The main reference of this artwork comes from 1970s style editorial design, with its minimal and optical shapes. The typeface used for the cover design is Apfel Grotezk, an open-source font designed by Colllettivo type foundry, Italy.

Contents

LIST OF PAPERS AND ABSTRACTS

The abstracts included here are the original abstracts of the published papers directly relevant to this thesis. In the introduction the abstracts have been re-written in order to holistically and organically weave the threads that run through this thesis.

PART I

Umbrello, Steven, 2021. "Coupling Levels of Abstraction in Understanding Meaningful Human Control of Autonomous Weapons: A Two-Tiered Approach." *Ethics and Information Technology*.

Abstract The international debate on the ethics and legality of autonomous weapon systems (AWS), along with the call for a ban, primarily focus on the nebulous concept of fully autonomous AWS. These are AWS capable of target selection and engagement absent human supervision or control. This paper argues that such a conception of autonomy is divorced from both military planning and decision-making operations; it also ignores the design requirements that govern AWS engineering and the subsequent tracking and tracing of moral responsibility. To show how military operations can be coupled with design ethics, this paper marries two different kinds of meaningful human control (MHC) termed levels of abstraction. Under this two-tiered understanding of MHC, the contentious notion of 'full' autonomy becomes unproblematic.

Umbrello, Steven, 2020. "Meaningful Human Control Over Smart Home Systems." *HUMANA.MENTE Journal of Philosophical Studies*, 13(37), 40-65.

Abstract The last decade has witnessed the mass distribution and adoption of smart home systems and devices powered by artificial intelligence systems ranging from household appliances like fridges and toasters to more background systems such as air and water quality controllers. The pervasiveness of these sociotechnical systems makes analysing their ethical implications necessary during the design phases of these devices to ensure not only sociotechnical resilience, but to design them for human values in mind and thus preserve meaningful human control over them. This paper engages in a conceptual investigations of how meaningful human control over smart home devices can be attained through design. The value sensitive design (VSD) approach is proposed as a way of attaining this level of control. In the proposed framework, values are identified and defined, stakeholder groups are investigated and brought into the design process and the technical constraints of the technologies in question are considered. The paper concludes with some initial examples that illustrate a more adoptable way forward for both ethicists and engineers of smart home devices.

PART II

Umbrello, Steven; van de Poel, Ibo, 2021. "Mapping Value Sensitive Design onto AI for Social Good Principles." *AI and Ethics*.

Abstract Value Sensitive Design (VSD) is an established method for integrating values in technical design. It has been applied to different technologies and recently also to artificial intelligence (AI). We argue that AI poses a number of specific challenges to VSD that require a somewhat adapted VSD approach. In particular, machine learning (ML) poses two challenges to VSD. First, it may opaque (to humans) how an AI systems has learned certain things, which requires attention for such values as transparency, explainability and accountability. Second, ML may lead to AI systems adapting themselves in such ways that they 'disembody' the values that have been embodied in them. In order to address these, we propose a threefold adapted VSD approach: 1) integrating the AI4SG principles in VSD as design norms from which more specific design requirements can be derived, 2) distinguishing between values to be promoted by the design and values to be respected by the design in order to ensure that the resulting design does not only do no harm but also contributes to doing good, and 3) extending the VSD process to encompass the whole life cycle of an AI technology in order to be able to monitor unintended value consequences and to redesign the technology if necessary. We illustrate the new VSD for AI approach with an example use case of a particular SARS-CoV-2 contact-tracing app.

1 Introduction

Power is information and information, power. Our current global epoch can arguably be defined by the exponential ability to compute information; thus, computers have ubiquitously ingrained themselves in every aspect of our quotidian existence, from the major to the banal. Notions of personhood, human essence, dignity, and the meaning of life have been brought under both scholarly and public scrutiny as these technologies shift traditionally held notions of what it means to be human in the age of artificial intelligence. Among others, the social, ethical, legal, and cultural issues regarding these technologies have therefore been the subject of intense scholarly debate and conversation in determining the current and future design and deployment of these technologies to ensure that they are beneficial to humanity and do not cripple human flourishing (Bostrom, 2014; Brynjolfsson & McAfee, 2014).

More poignantly, the development of these information technologies within the military sphere has garnered significant attention, as their implementation as constructs capable of force – a traditionally human-human affair – come with new ethical and legal issues surrounding machine autonomy, human dignity, and just war theory, among others. It also becomes a deeply personal affair, as the abdication of the capacity to select and kill targets without human interference proves instinctively controversial. The ethical and legal norms that have been historically developed to adjudicate the justified use of violence and how to deal with recalcitrant force likewise become the center of debate as autonomous weapons systems (AWSs) have been spotted on the developmental horizon. The use of armed drones – unmanned aerial vehicles (UAVs) – can arguably be characterised as the beginning of the technological divide separating humans from the direct use of force, although humanity retains the ultimate kill command over the release of such force.

The next step in the proliferation of automation consists in (fully) AWSs, in which the divide – both physical and psychological – appears to be absolute regarding human operators and the robots themselves, and the target selection and payload release are done without human confirmation or intervention (Docherty, 2012). It is the aim of this dissertation to provide some guidance to both the specialist reader as well as the international community at large in sober response to the tensions that have arisen from AWSs. In doing so, the concept of "autonomy" is brought to the fore, raising the central question as to what exactly constitutes autonomy and if full autonomy can and should be designed in AWSs. To this end, this dissertation takes the concept of meaningful human control (MHC) as its main conceptual and philosophical framework in tackling these issues. This concept, arising within the heated discussions on AWSs, has traditionally come to mean meeting the minimum

sufficient condition of having personnel "in/on the loop" who can be held accountable, thereby avoiding a "responsibility gap" that may emerge with the full autonomy of systems (Santoni de Sio and Mecacci, 2021). Discussions that took place in Geneva in 2014 and 2015 regarding the regulation of AWSs has led to a more holistic concept of MHC. Although the term "MHC" is often used haphasardly when speaking about AWSs, it provides a basis which ban supporters can sink their teeth into, that is, a partial ban on *fully* AWSs, thereby escaping the seeming paradox of having human control over a fully autonomous system.

In both public and scholarly debates, AWSs have been subject to three central ethical criticisms: (1) *fait accompli*, autonomous systems will not have the capacity to distinguish and execute the sophisticated practical and moral categories necessary for the level of compliance demanded by the laws of armed conflict (Guarini & Bello, 2012; N. E. Sharkey, 2008). These laws require compliance to satisfy *jus in bello*, by meeting the minimum necessary conditions for distinguishing between combatants and non-combatants, such that the proportionality in the use of force is similarly distinguished and that such use of force against non-military targets is not disproportional to the desired military outcome (Heyns, 2013). To this end, it can be clarified *prima facie* that the abdication of the use of force to (fully) autonomous systems raises significant legal and ethical issues. (2) The abdication of the use of force that may ultimately serve lethal ends is *mala in se*, meaning that their deployment is fundamentally immoral because it raises ethical concerns regarding human rights and, more critically, what it means to preserve human dignity – and dying a dignified death – in contexts such as wars (Sparrow, 2016; Wallach, 2013). (3) Either through maleficent use, design, deployment, or technical/human error, (fully) AWSs will create a liability vacuum, in which the responsibility gap between failure/misuse and attribution of responsibly can become severed (Chamayou, 2015; Heyns, 2013).

For the above reasons, the literature and debate has spawned concepts and arguments supporting the necessity of the principle of meaningful human control. The specialist non-profit organisation Article 36, which focuses on reducing harm caused by weapons, defines MHC over AWSs as follows:

> [It is] required in every individual attack. Sufficient human control over the use of weapons, and their effects, is essential to ensuring that the use of a weapon is morally justifiable and legal. Such control is also required for accountability over the consequences of the use of force. Critical aspects of human control broadly relate to:
>
> - The pre-programmed target parameters, the weapon's sensor-mechanism and the algorithms used to match sensor-input to target parameters.

- The geographic area within which and the time during which the weapon system operates independently of human control.

Similarly, states must understand:

- the process by which a system identifies individual target objects, and
- understand the context in space and time where an attack can take place.

(Article 36, 2015)

The principle was introduced to provide a more holistic and thus meaningful form of control over AWSs, rather than the difficult-to-define and often self-undermining concept of what exactly constitutes having humans "in/on-the-loop" (Crootof, 2016; Roff & Moyes, 2016). Thus, MHC appears to permit issues regarding human dignity – what can be interpreted in certain international contexts as being essential to understanding human rights – to be foundational in considerations regarding the legality of AWSs.

However, the difficulty that policymakers currently face is detailing the exact nature of evaluating the quality of control that can be deemed to be meaningful, the level of autonomy in systems and networks thereof that can be technically encompassed by such a definition, and the design specifications that can be adopted to operationalise such concepts in practice.

In their paper *Meaningful Human Control over Autonomous Systems: A Philosophical Account*, philosophy of technology and ethics scholars Filippo Santoni de Sio and Jeroen van den Hoven provide a novel and more philosophically nuanced account of how to conceptualise MHC as well as preliminary suggestions for operationalising such a concept in design. In exploring the concept of MHC, this thesis threads the various conceptions of MHC as presented in the literature, focusing primarily on Santoni de Sio and van den Hoven's conception, which is arguably more philosophically nuanced and robust. In doing so, it is my aim to deconstruct the philosophical underpinnings that constitute their understanding of autonomy and the role it plays in satisfying the conditions critical to MHC possession. If successful, the thesis will demonstrate the conceptual feasibility of satisfying a robust principle of MHC that can be applied to *fully* AWSs (and fully autonomous systems in general), as well as the case in which the conditions of MHC can be *buttressed* through an increase in systems autonomy if designed appropriately.

In addition, this thesis aims to explore the operationalisation of MHC in a responsible manner, thereby bringing it in line with the general objectives of responsible research and innovation (RRI) that are foundational to multinational parties such as the EU and the UN, with the aims of developing technologies and techniques that are sustainable and compliant with the key values of stakeholders (Groves, 2017; United Nations, 2018; van den Hoven & Jacob, 2013). To this end, the value-sensitive

design (VSD) methodology is adopted as the principled and philosophically grounded design framework for the operationalisation of MHC over (fully) AWSs.

The articles referenced in this dissertation – which have previously been published elsewhere – jointly build the foundation of the various concepts that I explore, both from a philosophical and a conceptual perspective. More specifically, the definition of the concept of MHC, latent issues with using existent conceptualisations, as well as how the VSD approach requires modification in light of some technical issues that have emerged from typically opaque artificial intelligence (AI) systems are explored. Although many of the papers explicitly mention and discuss these approaches and concepts with regards to applying them to discrete technologies such as general AI, there is no specific or exclusive focus on this context of application. This rests on the notion that, at the abstract level of theory formation and philosophical reflection on autonomy, meaningful control and technological design – on which this thesis focuses, as explained later in this dissertation – this difference in context and discrete application is non-essential.

In this introduction, I discuss various elements in order to place the proceeding sections and chapters in a broader conceptual prospect and delineate the veins that run through them. First, I discuss the motivation behind this specific project, the challenges encountered in such an endeavour, as well as the potential boons that await should the reader deem them sufficient in meeting their objectives (§1.1). Second, I outline the state of the art in the research on MHC, autonomy, and the VSD (§1.2). Third, I explain the central guiding questions that drive this dissertation and consider the implications of "operationalising" MHC on (*fully*) AWS (§1.3). Fourth, I raise issues surrounding autonomy in the military context, which adds further nuances to the underlying philosophical structure of MHC (§1.4). Fifth, I present a reading guide with a preview of the various chapters (§1.5). Finally, I conclude with some potential suggestions for fruitful research projects (§1.6). I assume that the reader of this doctoral thesis is familiar with the concepts of MHC; otherwise, I suggest deferring first to Annex I, which provides the necessary background on the topics covered later in this discussion.

1.1 Project Background: Developments and Methods

As with academic papers published in peer-reviewed journals, it is common practice to justify the merits of each piece of research for publication by determining the challenges that are currently being dealt with in the scholarship and how the article in question aims at addressing such a research gap. However, as has become common practice in ethnography and sociocultural anthropology, it is

advisable to determine the influences on any one author and how such influences have consequently affected the work (see also 'About the Author'). To this end, I use the present section to outline the main practical and theoretical influences underlying this work.

Beginning with the practical side, the organisational and structural style of this dissertation is heavily influenced by the paper-based doctoral dissertation of Dr. Ilse Oosterlaken, who completed and published her thesis entitled *Taking a Capability Approach to Technology and Its Design – A Philosophical Exploration* at the Technical University in Delft, Netherlands on January 15, 2013. The table of contents as well as the organisation of sections in this dissertation mirrors much of hers; however, given the originality of this thesis and the difference in topic, there are also significant changes which reflect differing viewpoints.

Similarly, the two annexes that follow this introductory chapter are meant to serve as the traditionally labelled "literature review" that is commonly included in dissertations. The decision to relegate the literature review to annexes, aside from a similarity to Oosterlaken's layout, is a stylistic one; it arguably improves the flow of the dissertation and conveys its central philosophical point. Dividing the literature review into parts helps the reader to determine what they can extract from it for their own research, as well as satisfies traditional academic norms of inclusion. How the literature review is conducted, however, differs starkly from Oosterlaken's methods, given that it is based on several contemporary approaches to conducting a literature review, primarily the methodology outlined in Justus Randolph's article, *A Guide to Writing the Dissertation Literature Review* (Randolph, 2009).[1]

With regards to its theoretical underpinnings, this dissertation can be categorised as building on the foundations laid down by the philosophy of technology in general, which has shifted away from the purely instrumental view of technology as neutral tools or artifacts adopted by humans. The shift away from this instrumental view of technology and towards an *interactive* one has been a fundamental stepping stone in what has been called the "design turn in applied ethics" (van den Hoven, 2017). In this view, technology is considered to be fundamentally value-laden and in a constant, co-constitutive relationship with stakeholders. Because technologies are laden with values, the question of why and how we design technologies to embody these values becomes of critical importance if such technologies are to benefit the stakeholder communities involved.

[1]Several other contemporary approaches to conducting and writing a literature review were considered, including the *integrative literature review* formulated by Richard Torraco (Torraco, 2016), Chris Hart's imaginative critical realism in mapping information (Hart, 2018), as well as the survey of systemic approaches method introduced by Andrew Booth, Anthea Sutton, and Diana Papaioannou (2016). Although the surveyed approaches all share common ground, the approach described by Randolph was ultimately chosen for its comprehensiveness and succinctness.

This broader trend of conceptualising technologies as interactional has led to the more specific concept of responsible innovation (RI),[2] which considers the ethical impacts that technologies and their design can have on societies as well as how to mitigate technological risks while engaging in ethically-driven design. Various design approaches that take the value-laden quality of technology as fundamental, such as the VSD, have been proposed as a means of attaining the objectives central to RI (Friedman & Hendry, 2019).

It was during my time at IEET (Institute for Ethics and Emerging Technologies) and GCRI (Global Catastrophic Risk Institute) that my cross-examination through various theories and principles ranging from molecular nanotechnology, artificial intelligence (including AGI/ASI[3] issues), existential risk theories, as well as posthumanist and transhumanist philosophies took place. While concurrently reading both science and technology studies (STS) at York University and ethics at the University of Edinburgh, those influences contaminated how I viewed ethics in technology. Value sensitive design has always been my primary nexus of research, exploring the strengths and areas for improvement within the approach. Naturally, my various scholarly backgrounds influence the means through which I address those challenges. Working with Seth Baum, Executive Director at GCRI, I conducted my first real research project, which culminated in a published paper in the journal *Futures* entitled *Evaluating Future Nanotechnology: The Net Societal Impacts of Atomically Precise Manufacturing* (2018). In it, we applied a consequentialist calculus to the net benefits and risks of atomically precise manufacturing in various domains spanning social, military, and environmental spheres (Umbrello & Baum, 2018). However, practical ethics as it concerns real people, in relation to technologies, resists being explained by the oppressive reduction of human values to economic ones that are central to the cost-benefit calculations of consequentialist and utilitarian approaches. Qualities such as beauty, calmness, love, and empathy, among others, can hardly be translated in any meaningful way by conceptualisations of ethics, neither utilitarian, consequentialist, nor Kantian.

To this end, my research drove me towards continental approaches to ethics, including the postphenomenology of technology as well as the posthumanist philosophies that seek to extricate themselves from the often sterile understandings of certain Enlightenment and humanist philosophies (i.e., *post*humanism). This avenue led to research on moral imagination theory, whose main proponents

[2]This concept, although not an old one, has already been established as a central concept in technology and research innovation within policy platforms and ethical guidelines by both private and public organisations, most notably the UN Sustainable Development Goals as well as the EU's goal for sustainable and responsible RRI in the Horizon 2020 objectives.

[3]AGI = artificial general intelligence; ASI = artificial superintelligence.

include philosopher Mark Johnson and cognitive scientist George Lakoff. The culmination of this approach led to a more holistic understanding of how human morality functions at real-world levels, rather than the narrow prototypical cases common to philosophical discussion (e.g., trolley dilemmas). This resulted in a published paper (2020) on the topic entitled *Imaginative Value Sensitive Design: Using Moral Imagination Theory to Inform Responsible Technology Design* (Umbrello, 2020a). The article aims to inform the VSD approach such that it would be sensitive to a more authentic understanding of human morality as informed by the cognitive sciences (i.e., moral imagination theory), and thus of human values (i.e., valuation), in how technologies are to be designed responsibly. A core position in this research project is that the meaning of autonomy, as understood in the literature on AWSs and MHC, does not necessarily reflect the technical and operational meaning of the concept as it pertains to AWSs within the military domain. If such is the case, the MHC of AWSs must be revised if RI is to be achieved in any meaningful sense. The VSD methodology has been proposed for this purpose, but it too must be revised if it is to meet the unique challenges posed by machine learning and artificial neural network-based systems which are proposed to be the main driving systems of (fully)AWS. To this end, this dissertation is a merging of *praxis* – that is, more poignantly, a contamination of systems thinking and engineering – and the applied ethics of analytic philosophy that has characterised the "design turn" (van de Hoven, 2017).

Having outlined some of the theoretical underpinnings of this project, two potential, albeit non exhaustive, questions may arise in the reader's mind: (1) Why this transdisciplinary approach – that is, what is gained by contaminating theories on MHC and VSD with more abstract approaches to technologies such as systems thinking and systems engineering? (2) Why choose the VSD as the approach for attaining the MHC of AWSs? In cursory response to the former, there is both a scientific and a conceptual gap between the theories developed during the Enlightenment on the nature of the human mind and, consequentially, its moral and autonomic faculties. These theories, like technologies, function as scaffolds that support as well as constrain and narrow to some extent the theories that proceed them, propagating certain discriminations and prejudices regarding norms and values. Intuitively, then, a re-evaluation of how the theories founded on such approaches and understandings becomes necessary in light of recent advances in the cognitive sciences that present alternative empirical explanations of how the human brain functions. The associated implications may further divide how we apply terms such as autonomy, responsibility, and moral agency to humans, and thus to autonomous systems such as AWSs. Likewise, VSD is chosen as the preferred approach for attuning this post-Enlightenment reconstruction of MHC, as it is a principled method of designing technologies,

one that is founded on the interactional perspective on human-technology relations as well as adapted to more situated and grounded understanding of human values, rendering itself sensitive to how humans *actually* engage in moral decision-making and valuation. Similarly, the approach has garnered the interest of multiple funding bodies by virtue of its potency in providing a means of achieving RRI. For example, in 2018, the European Research Council awarded a 2.5-million-euro ERC Advanced Grant to Delft Design for Values researcher Ibo van de Poel, who adopted the VSD as one of the primary theoretical approaches to technology design for stakeholder values. Similarly, Oosterlaken, Grimshaw, and Janssen (2009) received a grant of 550,000 euros from the Netherlands Organisation for Scientific Research (NWO) as part of their grant program, "Responsible Innovation," to which the VSD approach was instrumental (TU Delft, 2012).

Overall, the stakes are high; AWSs remain on the horizon, despite various multination organisations calling for a ban (such as the ICRAC[4] and the Campaign to Ban Kill Robots). Whether or not a ban will be effective is beyond the scope of this dissertation; although it may overlook major players who adhere to international treaties and agreements, there is nonetheless a tactical advantage in having possession of such arms, and thereby the incentive to develop them. It is my hope that the research conducted here can provide a "middle path," viz., design requirements that account for values that are important to all the stakeholders involved. If a sufficiently robust definition of MHC can be achieved for (fully) AWSs, then, by definition, a ban need not be the center of concern; rather, its pursuit would come at the opportunity cost of directing attention to the operationalisation of MHC in those AWSs.

1.2 Research on MHC and Technology: The *Mise en Scène*

The concept of MHC, which originated within the AWS debate in 2014 (Article 36, 2014), has attracted global attention and support from both nation states and ban advocates, as well as those who criticise the arguments that these advocates have proposed (Biontino, 2016). More than two dozen states are in support of a ban on (fully) AWSs, all of which support the principle of MHC as a necessary requirement for lawful AWSs to be deployed, so as to ensure that human control is never

[4]The International Committee for Robot Arms Control (ICRAC) "is a non-governmental organization (NGO). We are an international committee of experts in robotics technology, artificial intelligence, robot ethics, international relations, international security, arms control, international humanitarian law, human rights law, and public campaigns, concerned about the pressing dangers that military robots pose to peace and international security and to civilians in war" (ICRAC, n.d.).

downplayed in the context of AWS design and deployment (Sauer, 2016; Senear, 2018). To this end, it has been proposed either that MHC must be integrated into some existing internationally binding norm applicable to all states, making such a statute easier to ratify, or that a *de nuovo* norm must be synthesised (Asaro, 2016; Morley, 2015).

Regardless of which route is followed, the fundamental challenge that must be addressed is determining what exactly constitutes and satisfies a principle of MHC in AWSs. Each state may interpret human control in a different way. Noel Sharkey, a strong proponent of a ban on (fully) AWSs, distinguishes five levels of human supervisory control over such systems:

1. A human engages with and selects a target and initiates any attacks.
2. The program suggests alternative targets, and a human chooses which one(s) to attack.
3. The program selects a target, and a human must approve it before the attack.
4. The program selects a target, and a human has a limited amount of time to veto it.
5. The program selects a target and initiates the attack without human involvement.

(N. Sharkey, 2014)

A state might interpret MHC as requiring the lowest levels, 1–3, to be true, whereby humans have final executive authority over self-chosen or system-chosen targets. This is a positive interpretation of human control and is commonly referred to as the human being "in the loop" (Nash, 2015). Similarly, states can interpret MHC as being satisfied by level 4, in which a human has the time to veto the chosen target of an AWS, constituting a "human-on-the-loop" paradigm (Nahavandi, 2017). Level 5 is the level of autonomy – and thus a lack of human supervisory control – that Sharkey, and ban proponents in general, are adverse to; i.e., full autonomy whereby the target is chosen and engaged with without any human involvement in the process (Sauer, 2016; N. Sharkey, 2014). However, the "human-off-the-loop" paradigm described in level 5 has been considered grounds for MHC, given that the design of the program making targeting decisions and executing those decisions lies in the hands of the programmers and system designers themselves (Carpenter, 2014; Heins, 2018).

Despite a surge in the appropriation of the term "MHC" and the various modalities that entities have defined it as, the arguably most nuanced and philosophically grounded approach to explicating what it can consist of is provided by Santoni de Sio and van den Hoven (2018), as mentioned above. Their fresh view on what constitutes MHC (discussed in detail in Chapter 3 and briefly in Annex I) has been appropriated as the theoretical approach to the ethical inquiry and control of novel technologies beyond the realm of AWSs. For example, the MHC that they propose has already been adopted as a way to understand responsibility and liability in the case of autonomous vehicle platooning, in which (semi)autonomous vehicles and human operators work in conjunction with one another, despite levels

of autonomy that would normally muddy the waters in liability attribution (Calvert, Mecacci, Heikoop, & de Sio, 2018). I myself have recently published on the application of their version of MHC to smart home technologies, specifically smart personal assistants such as Google Home and Amazon Alexa (Umbrello, 2020b).

Most of the academic work within the field of MHC on AWSs and autonomous systems in general has been conducted only within the last few years. Much of the discussion surrounding Santoni de Sio and van den Hoven's MHC are outlined in both Calvert et al. (2018) and Umbrello (2020). The contents of these two papers, along with the original paper on this version of MHC (2018), are detailed in Chapter 3.

Nonetheless, much of the work on MHC has been less about defining what constitutes it, and more about the means through which such a constitutional entity, if definable, can form a defensible and enforceable international program across both a ban as well as a regulation of permissable forms of (semi-)AWS. To this end, this disseration is comparatively unique in that it builds on the last few years of scholarship on MHC, aiming to delve into and critique the typically preseumed philosophical substratum that lies at the foundation of the MHC discourse, and to construct a more holistic definition of MHC that can, if successfully demonstrated, be applied to certain forms of (fully) AWS. The geopolitical boons of such an enterprise need not be stated. Likewise, formal investigation as to how such a revision of MHC can be operationalised via VSD is comparatively unique to past (albeit still relatively recent) applications of VSD to Santoni de Sio and van den Hoven's MHC (Santoni de Sio & van den Hoven, 2018; Umbrello, 2020b).

1.3 Main Guiding Questions: Exploring Autonomy and Operationalising MHC in Terms of VSD

The preceding section aimed to provide a cursory overview of the current landscape of MHC while also briefly touching on some initial gaps in the research that warrant more attention. Given that these areas of research – MHC, AWSs, and VSD – are relatively recent subjects of scholarship and public debate, many of the potential areas addressing the questions that arise have yet to be formulated, and the field of applied ethics in technology is far from being saturated. My published research thus far has been primarily based on the question: *how can we design transformative technologies that cater to stakeholder values, and what design methodologies can we adopt to achieve those ends?* Value-sensitive design has been the primary and central approach in my research, albeit not without its own

philosophical issues (discussed in Annex II). Given the marked global increase in both scholarly and public discussions on artificial intelligence (AI) systems, the *design question* becomes of central importance to the philosophical debate on how we can guide the development of AI systems towards beneficial ends, however the concept of "beneficial" may be construed.

Because AI systems are foundational to the heated debate on the socioethical and legal issues that surround AWS development and deployment, the design question similarly delves into the following discussion: *if MHC can be conceptually achieved for either or both semi- and fully AWSs, what design approach can be adopted to best implement MHC?* Santoni de Sio and van den Hoven (2018) briefly mention VSD as a potential approach for implementing MHC in autonomous systems:

> Responsible Innovation and Value-Sensitive Design research focuses on the need to embed and express the relevant values into the technical and socio-technical systems (Friedman and Kahn, 2003; van den Hoven, 2007, 2013). From this perspective, the question to be addressed is how to design technical and socio-technical systems which in accordance with the account of meaningful human control we have here presented. Based on our analysis of meaningful human control, we propose the following two general design guidelines, and we briefly show how these can be applied outside the military context, by looking at the case study of automated driving systems (aka. "autonomous vehicles," "self-driving cars," "driverless cars"; Santoni de Sio and van den Hoven, 2018, p. 11).

Although their mention of how the values central to MHC can be cast as design requirements that play a crucial role in the operationalisation of MHC in VSD, this point is cursory. The philosophical exploration central to this dissertation thus follows from an ambiguity in the general literature on MHC and AWSs that is nonetheless central to any real progress towards the RI of such AWSs or even towards a ban. That is the concept of autonomy, and what technically constitutes autonomy in (A)WSs. In exploring the concept of autonomy, I draw on the concepts of systems thinking (and systems engineering) to underline the co-variance and co-constitution between human and machine autonomy that is fundamental to the understanding of either, particularly within the context of military operations planning and deployment (see Chapter 2).

Originally used by Bell Telephone Laboratories, systems engineering has been an increasingly popular approach to engineering technologies, one that has been on the rise since its general conception in the 1940s (Schlager, 1956). American engineer Simon Ramo popularised the concept from the 1950s onwards, defining it as "a branch of engineering which concentrates on the design and application of the whole as distinct from the parts, looking at a problem in its entirety, taking account of all the facets and all the variables and linking the social to the technological" (Hambleton, 2005, p. 10). The overall aim of this approach to artifact design is to understanding complexity holistically with technologies that

form parts of larger systems and are themselves systems (i.e., constituted of various heterogenous nodes). This socio-technical relationship has been fundamental to the sociology, anthropology, and philosophy of technology that underlies the STS approach to technological analysis, substantiating the inextricable link between nontechnical and technical entities. The inseparability of the various facets calls into question concepts such as autonomy, viewed as a discrete concept extracted from the sociotechnicity of the systems in question, AWSs or otherwise.

The subsequent goal, then, is to argue for the necessity of a more ontologically grounded theory of autonomy as it pertains to the military-industrial complex in achieving a meaningful floor for MHC in terms of AWSs. More saliently, for any MHC concept to be effective, it must map onto an ontologically secure ground regarding the meaning of autonomy. In doing so, how *full* autonomy is construed shows that MHC can be achieved through an increase in certain forms of human-machine autonomies, and that such technical requirements can be achieved through a VSD approach.

1.4 Human and Machine Autonomies: Outlining the Divide

In bringing to bear the essential guiding question at the root of this dissertation – *how can we design transformative technologies that cater to stakeholder values, and what design methodologies can we adopt to achieve those ends?* – many other philosophical issues worth investigating arise, particularly when highly controversial discrete technologies such as AWSs become the topic of consideration. Given the structure of this project, instead of providing an arid list of relevant issues that merit close consideration, I opt to simply refer to the proceeding chapters that are dedicated to clarifying them. That being said, one of the most interesting and central questions of this research project is the following: *what is the nature of autonomy as it relates to humans and machines in the military domain, and how does an understanding of human-machine autonomies and relationships change the meaning of MHC?* There is a considerable amount of information packed into this question. Bringing the concept of autonomy into question requires an intimate understanding of the literature across various fields that appropriate the term, including psychology, moral and political philosophy, and engineering, among others.

Of course, no comprehensive view is agreed upon by all in terms of what it means for something, whether human or non-human, to be autonomous. For example, Sartor and Omicini (2016) distinguish the autonomy of AWSs as consisting of three "dimensions": (1) independence, (2) cognitive skills, and (3) cognitive-behavioral architecture (Sartor & Omicini, 2016). This dissertation does not claim to provide such a comprehensive definition, lest it meet Icarus's fate. What it does aim to do,

however, is direct how we interpret autonomy when we speak about military operations (since it is the domain of interest here), and how this warrants consideration during the design phases (e.g., VSD) of AWSs if MHC is to be achieved. Although the introductory chapter does not provide the medium for discussing this in any detail, it bears noting that the theoretical underpinning adopts the more interactional and systemic approach at understanding the military-industrial complex in order to better grasp what autonomy can mean. In doing so, it aims to bridge the severing of praxis so as to inform the more analytical applied ethics of design.

In light of the above considerations, this doctoral dissertation can be read as being differentiated into two distinct philosophical parts:

- Part I, divided into three chapters, is markedly ontological. That is, it aims to show how *full* autonomy is not *mala in se*, but rather that increased autonomy can actually augment the ability to attain MHC in *certain* types of AWSs. To this end, systems thinking is used as the concrete landscape upon which a more ontologically grounded understanding of MHC can be framed.
- Part II, divided into four chapters, is markedly ethical. Through the lens of designing *for* values, it explores how VSD can be used as the approach to design AWSs so as to attain MHC (as defined under the systemic understanding of autonomy proposed in Part I).

The ontological explorations of autonomy as well as systems thinking provide the general philosophical basis upon which the latter part of the dissertation can take the practical, applied steps. Taken holistically, the chapters of this dissertation aim to argue that a systems view of the sociotechnical relations between humans and AWSs within the context of military operations planning allow for an understanding of *full* autonomy that can be achieved under MHC via VSD. However, the latter part of the thesis, in which VSD becomes the emphasised paradigm, is not taken *prima facie*, but rather brought under similar philosophical scrutiny as I have done in other articles, and discussed in greater depth in Chapter 5. The traditional conception within the VSD literature of the philosophical foundations and the process of valuation of stakeholder values in the design process is called into question (see Annex II). The work undertaken in Part I requires VSD to be sensitive to multiple levels of abstraction in design; more poignantly, VSD must be sensitive to the operational and organisational norms of the military-industrial complex (see Chapter 6) as well as employ full-lifecycle monitoring to avoid unforeseen (or unforeseeable) recalcitrance (see Chapters 6 and 7). Taken together, this dissertation aims to provide a more focused understanding of both MHC and VSD in practice.

1.5 Reading Guide and Paper Previews

In the previous subsection, I explained the main tensions underlying the aim of this thesis as well as the structure of the project itself, detailing the mains parts of the work as well as briefly summarising each of the chapters. Here, I outline the various papers that are included in some substantive form (fully or partially) in the dissertations. Unlike primarily paper-based dissertations (e.g., Oosterlaken, 2013), the published papers and chapters used sporadically throughout this work are used to support the arguments and aims that constitute the objectives of this dissertation rather than construct the dissertation itself.

What makes this particularly hybrid approach interesting is that it enables a wider audience to pick this work up and read the parts necessary or relevant to them without loss of fidelity. Because many of the papers included are directed at different audiences – viz., primarily to philosophers of technology or to engineers/designers – those chapters can be read as discrete works in and of themselves, even though they provide the medium of germination for the chapters that proceed them. This, of course, does not mean that the chapters primarily directed at one audience would not be of interest to others – this dissertation is a self-proclaimed trans-/interdisciplinary enterprise – but rather that they need not be read as such.

Chapter 3, for example, is a streamlined version of a paper originally published in the journal, *Humana.Mente*. This paper argues that VSD provides a strong design approach to framing and designing *for* MHC in smart home systems. Although such a paper has implications for how engineering practices are to be conducted, engineers may most likely be lost in the nuanced philosophical style. Meanwhile, philosophers of technology and theoretically oriented designers who are more familiar with various design approaches such as VSD may find more of value and interest. Designers and engineers who are more practiced-oriented, for instance, may be more inclined towards Part II, in which abstract values are demonstrated and more concretely translated into technical design requirements.

Paper Title	Published in	Target Audience	Possibly of Interest to
PART 1:			
Coupling Levels of Abstraction in Understanding Meaningful Human Control of Autonomous Weapons: A Two-Tiered Approach	*Ethics and Information Technology* (2021)	Policymakers	Philosophers of technology/ theoretically oriented designers

3. Meaningful Human Control Over Smart Home Systems: A Value-Sensitive Design Approach	*Humana.Mente Journal of Philosophical Studies* (2020)	Philosophers of technology	Policymakers
PART 2:			
Mapping AI for Social Good Principles onto Value-Sensitive Design	*AI and Ethics* (2021)	Systems engineers	Programmers/systems engineers/policymakers

Table 1. Individual papers included in this dissertation

Table 1 is intended to allow the reader to quickly navigate the included papers as well as orient their within the dissertation as a whole. The numbers to the left of the paper title represent the associated chapter which forms part or all of the paper. Where number are absent, the associated paper is used throughout the entire part of the thesis. This is a useful tool, since reading the dissertation as a whole, depending on the audience, can become repetitive, seeing as multiple papers detail some of the same conceptual tools, frameworks, and approaches – such as VSD, which is outlined in many of the included papers. However, unlike other paper-based dissertations, this project does not leave the articles in their original form. In order to increase readability and symbiosis between chapters, the styles of the introductions and conclusions of the included papers are changed, and much of the body of those works is dispersed among a large quantity of original work for this project. The original abstracts can be found in the section preceding the introduction. Following this paragraph, the reader can find the abstracts of the chapters containing the included papers, albeit slightly modified to render the transitions between the preceding and proceeding chapters more seamless.

PART I: A PHILOSOPHY OF SYSTEMS THINKING AND MEANINGFUL HUMAN CONTROL

Coupling Levels of Abstraction in Understanding Meaningful Human Control of Autonomous Weapons: A Two-Tiered Approach
Originally published in 2021 in the journal *Ethics and Information Technology:*

Chapter 2 – Systems Theory: An Ontology *for* Engineering

In order to bridge the levels of abstraction and thereby conceptualise a unified theory of MHC over AWSs, as well as to subsequently unify this conception of MHC with a design approach that is capable of designing *for* it (i.e., VSD), this chapter proposes systems thinking as the ontological substrata. The main reason for adopting this approach is that it (implicitly)

characterises the two levels of abstraction for understanding MHC. The operational level of control is characterised by a plurality of actors and networks that complicates but also constitutes how military operations are structured, planned, and conducted. Likewise, the design level of control is fundamentally built on the notion of tracking and tracing networks of systems and actors within both the use and the design histories of those systems. In addition, systems thinking is the theoretical framework from which systems engineering derives. It is essentially the practical and managerial implementation of a systems thinking ontology, whereas VSD exists as a sort of parallel approach to the systems thinking design methodology

Chapter 4 – Coupling Levels of Abstraction: A Two-Tiered Approach

The marriage of both levels of MHC (i.e., the operational and design levels) is demonstrated to be symbiotic with regards to MHC. Here, the argument is that military operations *always already* constrain the autonomy of any and all agents within the military-industrial complex as a function of the procedures that necessarily take place *a priori* to the deployment of force (i.e., the operational level). Close cooperation between institutions and infrastructures that constitute the military-industrial complex (e.g., the military, industry, government, and legislative norms) likewise form the *supraindividual* agent that can be said to be the possessor of MHC, if the design history can be *traced* and its behaviors can be *tracked* to the relevant moral agents (i.e., MIC). These two levels of abstraction warrant closer cooperation within the MIC so as to allow more accurate mapping of the moral intentions of the aforementioned agents onto AWSs that are being developed/deployed. The consequence here is that, if MHC obtains across both levels of control, then not only is autonomy *per se* not the problematic vector, but it can actually be increased, thereby increasing MHC.

Meaningful Human Control Over Smart Home Systems: A Value-Sensitive Design Approach

Originally published in 2020 in the journal *Humana.Mente Journal of Philosophical Studies:* 13 (37), 40–65.

Chapter 3 – Meaningful Human Control: Two Approaches

To couple the various levels of abstraction, this section builds on the literature review of Annex I, in which both Ekelhof and Santoni de Sio's works on MHC, among others, are explained. In

this chapter, the approaches presented in these papers are discussed, in addition to how we can begin to view those approaches as symbiotic in terms of their systems thinking affinities. The initial groundwork is then laid for understanding how they both complement each other without encumberment.

PART II: DESIGNING MEANINGFUL HUMAN CONTROL WITH VALUE-SENSITIVE DESIGN

Mapping AI for Social Good Principles onto Value-Sensitive Design
Originally published in 2021 in the journal *AI and Ethics*:

Chapter 5 – Value-Sensitive Design: Conceptual Challenges Posed by AI Systems

Value-sensitive design has been adopted as a principled approach to designing various existent as well as futuristic/transformative technologies. The VSD approach is fundamentally predicated on the interactional stance towards technology – or, more precisely, that societal and social factors co-construct and co-vary with technological artifacts. Part of the rationale behind this approach is that technologies embody values. However, AI systems that employ machine learning (ML) and/or artificial neural networks are often opaque, and thus the values that they may (dis)embody can be unforeseen or unforeseeable. This chapter discusses the different ways in which technologies embody values and how they fit within the larger systems thinking approach, as well as how to more saliently frame the embodiment of values *for* AI systems such as AWSs.

Chapter 6 – Adapting the VSD Approach

As ML systems (often) learn in ways that are opaque to humans, we need to pay attention to values such as transparency, explicability, and accountability. To address this issue, as well as the potential "disembodiment" of certain values over time, I propose a threefold, modified VSD approach: (1) integrating a known set of VSD principles (AI4SG) as design norms, from which more specific requirements can be derived; (2) distinguishing between values that are promoted and respected by the design to ensure outcomes that not only prevent disproportionate harm but also actively promote just war; and (3) extending the VSD process to encompass the whole

lifecycle of an AI technology, so as to monitor unintended value consequences and redesign as needed.

Chapter 7 – The AI4SG-VSD Design Process in Action: Multi-Tiered Design and Multi-Tiered MHC

The AI4SG-VSD approach described in the previous two chapters is employed with the AWS as the use case. In doing so, I outline the values to be promoted as much as possible (e.g., the LOACs), the (constraining) values to be respected as much as possible (e.g., the EU HLEG AI), as well as the AI4SG norms as a means for translating these abstract values into technical design requirements. The value hierarchy is chosen as the tool for illustrating how designers can begin to conceptualise this translation to design *for* values rather than *ex post facto*, ad hoc, or not at all. Likewise, I discuss how full-lifecycle monitoring and incremental deployment into an envelope of safe use to determine the emergent behaviours and consequent implicated values can be used to evaluate whether a system requires a redesign. In the event that this cannot be done, such types of systems should be considered *de facto*, or otherwise prohibited, given the associated risks of bypassing such an approach.

1.6 Conclusions

In summary of this introduction, it is worthwhile to note the importance of the explorations undertaken by this dissertation in the proceeding sections. Undoubtedly, exploring the notion of the MHC of AWSs comes with obvious sociopolitical and ethical boons. The definition of MHC, however, is another matter, as is the practical implementation of any meaningful conception of MHC. In the end, the latter question may prove to be the most difficult hurdle; but first, the question of what to design must be brought to the fore. Deciphering the notion of autonomy, given its indispensability to AWSs (it is, after all, the first letter of the acronym), is critical to understanding how AWSs function and tracking threads of accountability and liability, among other issues. Drawing on fundamental notions within systems thinking, military planning, and engineering, provides important conceptual tools and initial steps to understanding the network of causation and responsibility in establishing an ontologically grounded understanding of human-AWS relations.

The ultimate goal of this dissertation is to re-center life (viz. human, animal, and environmental, among others) as the object being designed *for*. That is, stakeholders – rather than the technology in

question – take center stage when discrete technologies are being considered. Seeing as AWSs are on the horizon, there is a growing anxiety that their development will run contrary to many human rights and values. These anxieties warrant worry, yet rather than succumb to technological determinism or instrumentalism, interacting with the technology early on and throughout the development programs of such systems can provide the middle way that is beneficial to all stakeholders. This, of course, is neither an admonition nor a statement in support *for* the development of AWSs and thereby their deployment for violent ends. I accept, however, that the "end of the war" is nowhere in sight, and that AWSs are more likely than not to be developed. This dissertation is my humble offering to the community currently engaged in the debate over a solution to design AWSs *for* stakeholder values and achieve more socially desirable outcomes.

References

Article 36. (2014). *Key areas for debate on autonomous weapons systems*. Geneva. http://www.article36.org/wp-content/uploads/2014/05/A36-CCW-May-2014.pdf

Article 36. (2015). *Killing by machine: Key issues for understanding meaningful human control*. Geneva. http://www.article36.org/weapons/autonomous-weapons/killing-by-machine-key-issues-for-understanding-meaningful-human-control/

Asaro, P. (2016). Jus nascendi, robotic weapons and the Martens Clause. In R. Calo, M. Froomkin, A. Michael, & I. Kerr (Eds), *Robot Law*. Edward Elgar Publishing.

Biontino, M. (2016). *Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapon Systems* (LAWS). United Nations.

Booth, A., Sutton, A., & Papaioannou, D. (2016). *Systematic approaches to a successful literature review* (Second ed.). SAGE Publications.

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

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W.W. Norton & Company.

Calvert, S. C., Mecacci, G., Heikoop, D. D., & de Sio, F. S. (2018). Full platoon control in truck platooning: A meaningful human control perspective. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. pp. 3320–3326. IEEE.

Carpenter, C. (2014). *Dynamics of debate at the experts meeting on autonomous weapons.* Duck of Minerva. Retrieved January 30, 2020, from https://duckofminerva.com/2014/05/dynamics-of-debate-at-the-experts-meeting-on-autonomous-weapons.html

Chamayou, G. (2015). *Drone theory*. Penguin Books UK. https://www.penguin.co.uk/books/268/268667/drone-theory/9780241970348.html

Crootof, R. (2016). A meaningful floor for "meaningful human control." *Temple International and Comparative Law Journal*, *30*, 53.

Delft University of Technology. (2012). *Technology and human development: Applying the capability approach of Sen and Nussbaum to technology, engineering and design.* Retrieved January 29, 2020, from https://www.tudelft.nl/io/onderzoek/research-labs/applied-labs/technology-and-human-development/

Docherty, B. (2012). *Losing humanity: The case against killer robots*. Human Rights Watch.

https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots
Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press.
Groves, C. (2017). Review of RRI tools project. *Journal of Responsible Innovation*, *4*(3), 371–374. https://doi.org/10.1080/23299460.2017.1359482
Guarini, M., & Bello, P. (2012). Robotic warfare: Some challenges in moving from noncivilian to civilian theaters. *Robot Ethics: The Ethical and Social Implications of Robotics*, *129*, 136.
Hambleton, K. (2005). *Conquering Complexity: Lessons for defence systems acquisition*. Stationery Office Books.
Hart, C. (2018). *Doing a literature review: Releasing the research imagination* (Second ed.). SAGE Publications.
Heins, J. C. (2018). *Letting Go of the Loop: Coming to Grips with Autonomous Decision-Making in Military Operations*. U.S. Naval War College.
Heyns, C. (2013). *Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions*, Pub. L. No. A/HRC/23/47. Human Rights Council, UN General Assembly. http://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf
ICRAC. (n.d.). About ICRAC. ICRAC. Retrieved January 29, 2020, from https://www.icrac.net/about-icrac/
Morley, J. (2015). Meaningful human control in weapons systems: A primer. *Arms Control Today*, *45*(4), 7.
Nahavandi, S. (2017). Trusted autonomy between humans and robots: Toward human-on-the-loop in robotics and autonomous systems. *IEEE Systems, Man, and Cybernetics Magazine*, *3*(1), 10–17.
Nash, T. (2015). *Remarks to the CCW on Autonomous Weapons Systems*. Geneva. http://www.article36.org/statements/701/
Oosterlaken, I. (2013). *Taking a capability approach to technology and its design: A philosophical exploration*. Delft University of Technology. https://repository.tudelft.nl/islandora/object/uuid%3Adf91501f-655f-4c92-803a-4e1340bcd29f
Randolph, J. (2009). A guide to writing the dissertation literature review. *Practical Assessment, Research, and Evaluation*, *14*(1), 13.
Roff, H. M., & Moyes, R. (2016). Meaningful human control, artificial intelligence and autonomous weapons. *Briefing Paper Prepared for the Informal Meeting of Experts on Lethal Au-Tonomous Weapons Systems.* UN Convention on Certain Conventional Weapons.
Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*. https://www.frontiersin.org/article/10.3389/frobt.2018.00015
Santini de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. Philosophy & Technology, 1-28.
Sartor, G., & Omicini, A. (2016). The autonomy of technological systems and responsibilities for their use. In N. Bhuta, S. Beck, R. Geiss, H.-Y. Liu, & C. Kress (Eds.), *Autonomous weapons systems: Law, ethics, policy* (pp. 39–74). Cambridge University Press. http://hdl.handle.net/1814/45234
Sauer, F. (2016). Stopping "killer robots": Why now is the time to ban autonomous weapons systems. *Arms Control Today.* Retrieved October 8, 2017, from https://www.armscontrol.org/ACT/2016_10/Features/Stopping-Killer-Robots-Why-Now-Is-the-Time-to-Ban-Autonomous-Weapons-Systems
Schlager, K. J. (1956). Systems engineering-key to modern development. *IRE Transactions on Engineering Management*, (3), 64–66.
Senear, M. (2018). "Killer robot" debates planned. *Arms Control Today*, *48*(1), 40.

Sharkey, N. (2014). Towards a principle for the human supervisory control of robot weapons. *Politica & Societa*, *3*(2), 305–324.
Sharkey, N. E. (2008). Grounds for discrimination: Autonomous robot weapons. *RUSI Defence Systems*, *11*(2), 86.
Sparrow, R. (2016). Robots and respect: Assessing the case against autonomous weapon systems. *Ethics & International Affairs*, *30*(1), 93–116.
Torraco, R. J. (2016). Writing integrative literature reviews: Using the past and present to explore the future. *Human Resource Development Review*, *15*(4), 404–428.
Umbrello, S. (2020a). Imaginative value sensitive design: Using moral imagination theory to inform responsible technology design. *Science and Engineering Ethics*, *26*(2), 575–595. https://doi.org/10.1007/s11948-019-00104-4
Umbrello, S. (2020b). Meaningful human control over smart home systems: A value sensitive design approach. *Humana.Mente Journal of Philosophical Studies*, *12*(37).
Umbrello, S., & Baum, S. D. (2018). Evaluating future nanotechnology: The net societal impacts of atomically precise manufacturing. *Futures*, *100*(June), 63–73. https://doi.org/10.1016/j.futures.2018.04.007
United Nations. (2018). *Transforming Our World: The 2030 Agenda for Sustainable Development*, Pub. L. No. A/RES/70/1 (2018). https://sustainabledevelopment.un.org/post2015/transformingourworld
van den Hoven, J. (2017). The design turn in applied ethics. In J. van den Hoven, S. Miller, & T. Pogge (Eds.), *Designing in Ethics*, pp. 11–31. Cambridge University Press. https://doi.org/10.1017/9780511844317
van den Hoven, J., & Jacob, K. (2013). *Options for Strengthening Responsible Research and Innovation*. https://doi.org/10.2777/46253
Wallach, W. (2013). Terminating the terminator: What to do about autonomous weapons. *Science Progress*, *29*.

Annex I: Meaningful Human Control – An Introduction

Introduction

> *If the problem is how to maintain meaningful human control of autonomous warfighting systems, no good solution presents itself* (Adams, 2001, 11)

The concept of 'meaningful human control' (MHC) originates from the discourse on autonomous weapons systems (AWS). It emphasises the notion that humans must remain in a position of control or oversight over the decision-making of a lethal system (Article 36, 2015; Morley, 2015). In other words, such types of systems should not be able to execute lethal action without human intervention. The above quote by scholar and political military strategist Thomas K. Adams belies the difficulty of formulating a practical solution (something that might constitute MHC) while also preserving the ever-increasing processing rates that accompany increased automation (Adams, 2001).

The literature on AWS that deals specifically with issues linked to human supervision and participation in the decision-making process can be divided into three interrelated categories. Each category involves an arguably distinct set of human capacities or features that are also privy to machines:

1. The assignment and abdication of responsibility, liability, and accountability (Allen & Wallach, 2014; Asaro, 2016; Scherer, 2015);
2. Humans as possessors of a discrete ability to make moral/ethical determinations, which is rooted in their empathic capacities (Asaro, 2009; Docherty, 2012);
3. The inability of machines to perform at certain levels or respond to certain situations that humans arguably can. At present, the system redundancy, error detection, and recovery architecture of machines cannot match the technical level of a comparable human equivalent in terms of function (Heyns, 2013).

These three categories of MHC are not limited to AWS *per se*, but apply to achieving MHC over autonomous systems in general. Recent scholarship has taken this challenge on by exploring the issue of achieving MHC over less directly lethal (yet still contentious) technologies such as autonomous vehicles (Calvert, Mecacci, Heikoop, & de Sio, 2018) and smart home systems (Umbrello, 2020). Scholarship has also addressed the general design and deployment of artificial intelligence (AI) systems that are socially beneficial and systemically resilient (Stephanidis et al., 2019). For readers unfamiliar with the literature on MHC, this introduction provides a robust account of various scholarly perspectives.

Technological innovation geared towards increased efficacy in war theatres has historically been the prerogative of militaries, garnering ever more attention at a global level today (Kania, 2017; Tucker, 2017). Similar attention has been paid, both in public and in academic debates within scholarly journals, to warfare innovations outside the military sphere (Altmann, 2005; Geiss, 2015; Walsh, 2015). At an international level, the UN Convention on Conventional Weapons was designed to address various issues regarding the legality and ethical development and use of AWS (Germany, 2014). One of the primary vectors of debate for this legal framework centred on what it means to exercise human control/supervision over these types of weapons. What current technological capabilities can support or constrain that type of control? Although there is no consensus on the particularities of what constitutes such control, there is convergence on a minimum standard of human engagement in the functioning of these types of systems (Crootof, 2016; Korpela, 2017). Aside from MHC, some other similar concepts have emerged (such as 'sufficient human control' and 'appropriate levels of human judgment'). However, the literature on MHC has proven most pervasive. Ekelhof (2019) provides a useful chart to capture the "recurring terms, themes, and elements in existing descriptions of human control standards" (Figure 1).

CNAS	**US DoD**	**Article 36**	**ICRAC**	**ICRC**
Human operators make informed, conscious decisions about the use of force.	The need for operators to make **informed and appropriate decisions** in engaging targets through readily understandable interface	Reference to timely **human judgment** and action.	There must be active **cognitive participation** in the attack and the ability to **perceive and react** to any change or unanticipated situations	Reference to **human intervention** in different stages (development, deployment, use).
Human operators have sufficient information to ensure the lawfulness of the action they are taking, given what they know about the target, the	Systems will be designed with appropriate human-machine interfaces and controls as well as appropriate safeties, antitamper mechanisms and	**Accurate information** for the user on the **outcome sought, the technology and the context of use.**	Reference to **deliberation on the nature of the target, its significance and likely incidental effects**. Also a reference to the need to have full	**Knowledge and accurate information** about the functioning of the **weapon system and the context** of its intended or expected use.

weapon, and the context for action.	**information assurance**		**contextual and situational awareness of target area**	
The weapon is designed and tested, and human operators are properly trained, to ensure effective control over the use of the weapon.	Need for rigorous **verification and validations, operational testing and evaluation** to ensure the systems function as anticipated.	Reference to need for **predictable, reliable and transparent technology** – that could be linked to **design** features	Reference to a means for the **rapid suspension or abortion** of the attack-that could be linked to design features	Reference to need for **predictability and reliability** of the weapon - that could be linked to **design** features.
Explicit reference to the need for sufficient information to ensure the lawfulness of the action is included in the element's description.	A reference to the need to employ systems **in accordance with the law** is made in the Directive but not as part of the standard itself.	**Accountability** to a certain standard. The requirement to make legal judgments is described in the broader analysis of the concept	**Necessity and appropriateness** of attack. Meeting the requirements of international law is reflected in broader statement as a driver.	**Accountability** for the functioning of the weapon system following its use. IHL compliance is considered a core driver of the concept.

Figure 1. Recurring terms, themes, and elements in existing descriptions of human control standards (Source: Ekelhof, 2019, 344)

Bolded terms show the relationships between each of the varying concepts. Although there are similarities between the concepts, there are also substantive differences. The primary philosophical underpinning that unites the various elements is the human-machine relationship. More specifically, it is the notion that there *is* a relationship between the human (operator or otherwise) and the autonomous system rather than pure independence (a point discussed in greater detail in Part I of this dissertation). The plurality of positions, as well as the various philosophical and/or legal motivations underlying these positions, contributes to ongoing difficulties in forging consensus on the conceptual and technical requirements that would meet necessary and sufficient conditions for MHC.

This difficulty is exacerbated by pressure on states to agree to legally binding tools ("The Campaign To Stop Killer Robots," n.d.) and political agreements (Germany/France, 2017), along with

other constructs, regarding their use. Pressure has increased in light of ongoing trends towards ever greater automation and the dehumanisation of warfare, wherein human combatants are removed from the war theatre (Marauhn, 2018). Regardless of the route that is taken, both the difficulty and prescience of having a converging theory of MHC lies in translating its more abstract concepts into a functional definition of *actual* military practices – there is difficulty moving from theory to practice, in other words. This is best illustrated by the International Committee of the Red Cross (ICRC), which has aimed to refocus discussion on speculative future weapons technologies by shifting attention to existing warfare systems in order to determine the relationships between humans and technology (ICRC, 2016). Knowledge of existing relationships can then be used as groundwork to inform discussions about more speculative systems.

As mentioned already, various approaches have been taken to address what constitutes MHC. For the sake of space and length, I do not discuss all of the literature on MHC. Rather, I focus on a selection of six papers (six approaches) that have tackled the issue from different approaches. This allows for a more comprehensive appreciation of the various perspectives on attaining MHC. The six approaches are as follows:

1. Preserving MHC through proper preparation and legitimate context for use, viz. through current NATO targeting procedures (Roorda, 2015);
2. Attaining MHC by having a human agent make "near-time decision[s]" in a AWS engagement (Asaro, 2012);
3. Preserving MHC through adequately training commanders in the deployment and function of AWS to ensure proper attribution of responsibility (Saxon, 2016);
4. Attaining MHC through apprising designers/programmers of their moral role in the architecture of AWS (Leveringhaus, 2016);
5. Attaining MHC through design requirements involving necessary conditions to *track* the relevant moral reasons for agent actions and *trace* the relevant lines of responsibility through design histories (Mecacci & de Sio, 2019; Santoni de Sio & van den Hoven, 2018);
6. Preserving MHC by distributing responsibility for decisions through the entirety of the military-industrial complex (Ekelhof, 2019).

2 Accounts of Meaningful Human Control

2.1 Targeting Procedures

Roorda (2015) locates the vector of MHC for AWS in the existing guidelines for NATO's targeting procedures. The author argues that AWS do not need to be able to distinguish or make proportionality decisions that human agents need to make as international humanitarian law (IHL) does not prescribe such a necessary condition. Rather, Roorda argues it is the 'effects' of attack decisions that must map onto relevant norms. Human operators and commanders are the nexus point upon which responsibility falls. He thus argues that an important factor for decision- making lies with those human agents. They are tasked with determining the appropriate context for use of any given system and its particular capabilities. NATO's existing targeting procedures provide this normative foundation, particularly given their incorporation of legal code, for the responsible deployment and use of arms including AWS. To that end, Roorda argues fully autonomous AWS may be used without *direct* human supervision – provided they can meet the normative requirements of NATO's targeting procedures as well as remain sensitive to informed decisions made by human operators about the proper context for deployment and use. Let us explore this in greater detail.

Roorda's argument rests on what he considers to be a privation in the debate on the autonomy of AWS: that these forms of arms are overly anthropomorphised, self-governing, and discrete (from human operators). Because of this, the focus on the legality of the weapons' ability to conform to normative moral requirements that has characterised the debate is fundamentally misplaced. Even if such weapons are capable of selecting and engaging targets without human selection and authorisation, they nonetheless remain within a larger human-machine network where the context for use is a highly relevant factor. Because *actual* military operations require planning and execution, types of weapons, their deployment, and the context for use are also governed by rule and procedures. It is during these phases that legal and ethical constraints are negotiated to ensure proper use of force, so it is here that the vector for MHC can be located for AWS.

Various normative frameworks already constrain assessments gathered and formulated during the planning phases of military operations. Here, NATO's targeting procedures combine these constraints to determine the appropriate and proportional use of force in an operation. These various legal and operational rules constitute very specific operational objectives that terminate in a single decision, which constrains whatever method of force is used regardless of its technological level of autonomy. Roorda (2015) sums up the decision-making procedure as follows:

> The doctrine defines joint targeting as: the process of determining the effects necessary to achieve the commander's goals (ICRC, 2018), identifying the actions necessary to create the desired effects based on the means available, selecting and prioritizing targets, and synchronising

> fires with other military capabilities, and then assessing their cumulative effectiveness and taking remedial action if necessary. (155)

Given that the decision-making process and final decision for operation are determined by humans, they implicate human responsibility for operational outcomes. Regardless of the types of systems used to carry out the final operational decision (even ones with autonomous targeting and engagement systems), responsibility for their use falls exclusively to humans, i.e., those who formulated the decision. This is because the operational process anterior to deployment constrains the set of appropriate targets *a priori*. For this reason, the autonomy of AWS co-varies with human operators. Systems are thus neither responsible for the formulation of such operational plans, nor their own place in the execution of those decisions. Similarly, the Laws of Armed Conflict (LOAC) do not specify the level at which compliance with legal norms is required. It would thus be absurd to require AWS to be compliant *per se*. Instead, compliance with the LOAC can be satisfied (as it normally is) during the operational decision-making process that determines targets, context for use, and the means of achieving objectives.

2.2 Near-Time Intervention

Here, we discuss the more technology-focused argument for attaining MHC derived by Asaro (2012). Alongside Jürgen Altmann, Noel Sharkey, and Rob Sparrow, Peter Asaro pioneered the position of the International Committee on Robot Arms Control (ICRAC) in favor of the prohibition of AWS. Given that the latter eliminate human judgment in the initiation of lethal force, they threaten to undermine bodies of international humanitarian law (IHL) and international human rights law (IHRL). Asaro defines AWS as "any automated system that can initiate lethal force without the specific, conscious, and deliberate decision of a human operator, controller, or supervisor" (2012, 694). In this, he acknowledges a nuanced point regarding what differentiates such systems from other independent weapons systems such as landmines or the auto-turret system: they are less 'weapons-as-tools' and more like a system that uses weapons or, more specifically, an autonomous weapons platform. Echoing what Merel Ekelhof would later write, Asaro notes that "autonomous weapon systems force us to think in terms of 'systems' that might encompass a great variety of configurations of sensors, information processing, and weapons deployment, and to focus on the process by which the use of force is initiated" (2012, 694). What Asaro describes captures the complexity of the technical systems that form AWS, and his point is not unimportant. The shift in perspective occurs not so much in terms of AWS-as-a-tool and even less as AWS-as-a-system. Instead, it positions AWS as within (or part of) a system

or network. This point forms the crux of the philosophical lining detailed in the first part of this disseration on reformulating a more systems-based notion of MHC.

In an effort to reduce the potential to undermine humanitarian or human rights law, Asaro proposes both minimum and necessary conditions that must exist for AWS to fall under MHC. Firstly, he describes what the US military designates as a 'kill chain' or, more aptly put, the process through which an order to execute is achieved: find, fix, track, target, engage, and assess. Asaro (2012) argues that having the so-called 'human on the loop' is the middle ground between (fully) AWS and the direct operation control of having a human-in-the-loop. This means the presence of a human at any single point in that six-step chain is a necessary but insufficient condition. For AWS to be under MHC, humans must be able to assess and verify the *target* and *engage* steps. According to Asaro, this is the defining characteristic of (fully) AWS. Abdication of these two steps to a process that is fully divorced from human involvement (i.e., purely in the hands of the machine) fails to meet the minimum standard for MHC. Failure then opens the floodgates for violation of international humanitarian and human rights law.

Consequently, a treaty defining the meaning of what constitutes AWS as well as their design, deployment, and use would be fundamentally predicated on compliance with international humanitarian and human rights law. Responsibility would be necessarily attributed to 'informed and trained' human operators making target and engagement decisions, all of which are currently delineated in current military practices governed by international treaties on the conduct of warfare. The ICRAC itself has formulated guidelines on the means through which target acquisition is deemed legitimite and in compliance with international humanitarian and human rights law.

2.3 Proper Commander Training

Echoing the potential for violations to humanitarian and human rights law observed by Asaro, Saxon (2016) argues that the general use of autonomous drones and AWS does not necessarily entail a responsibility gap in terms of attributing individual moral responsibility to a human in the kill chain. He reviews the literature on criminal responsibility to show the existing theories applicable to crimes committed through use of these weapon platforms (aerial autonomous drones, in his case). But he concedes that as advancements in these technologies augment, and as commanders abdicate more of their supervisory control, the issues of responsibility attribution described in criminal responsibility theories become more challenging. Still, he never acknowledges their inability to address such issues.

Saxon (2016) argues that compliance with international humanitarian law requires human supervision across four stages of military operations:

> (1) the procurement/acquisition stage, (2) the planning stage of the mission or attack when a human must choose which weapon system to employ (systems will vary across a range of autonomy) (echoed by Ekelhof in 2.5), (3) following the choice of an autonomous drone, a decision as to the level of human attention – if any – to assign to the system for the mission, but prior to the attack, and (4) specific inputs of human judgment – if necessary – to comply with international legal obligations and/or political interests immediately before, during, and after the attack. (18-19)

Moreover, the human supervisor must monitor continued legal compliance throughout stages 2 to 4. If crimes are committed through the use of these systems, the degree of autonomy present in such systems must be accounted for in any analysis of criminal liability. He then mentions that (fully) AWS may preclude *mens rea* entirely, strangely enough, which would sever individual responsibility for crimes committed.

Saxon locates the vector of responsibility in conventional criminal law, where crimes committed by a AWS must be found in the human operators or commanders who (whether through negligence or intent) fielded the AWS *contra legem*. Of course, the customary minimum necessary conditions of *habeas corpus* apply in regards to having sufficient evidence of such intent. Attribution of responsibility can even be assigned in a 'superior' way. This means a commander can be held personally responsible for criminal acts of omission rather than commission or direct intent, which are governed under direct responsibility. The finding holds true even for commander-subordinate complexities in the military hierarchy of criminal orders passed down (omission). Criminal acts of commission by a commander, such as knowingly deploying AWS in civilian-dense regions, can be used as evidence for the attribution of direct criminal responsibility. Technical measures during design can enable tracing lines of responsibility to support *mens rea* in terms of commands given (either directly from a commander or by way of subordinates) through various ledger systems within the AWS themselves.

This puts the ultimate responsibility for the use, deployment, and amelioration of potential malfunctions on commanders. Thus despite the ever-increasing independence, speed, and complexity of autonomous systems, proper training is needed. Training must include an effective means of shutting down systems when the first signs of potential recalcitrance emerge to ensure there is no risk of violating the laws of armed conflict – regardless of the economic costs of the system itself. MHC here means proper training for commanders so their decisions remain discretely within their domain throughout the planning and fielding stages.

2.4 The Moral Responsibility of Designers

Leveringhaus (2016) takes an approach similar to both that of Roorda (2015) and Santoni de Sio et al (2018, 2019). The former locates MHC within targeting procedures, while the latter locates it partly in relation to relevant designers/programmers. Leveringhaus explores the distinction between allocating moral responsibility for both semi- and fully AWS between drone pilots and programmers. Tackling the challenges that emerge from the allocation of moral responsibility, he argues these issues are best confronted with what he calls a 'Standard of Care Approach'. Leveringhaus predicates his analysis and application of the Standard of Care approach on three background considerations: automated targeting, moral responsibility, and Just War theory.

Automated Targeting

The notion of automated targeting as a strong disjunction (i.e., either there is automated targeting or there is not) is a fallacious one. Instead, Leveringhaus (2016, 169-170) distinguishes five different stages of the decision-making process (the so-called kill chain) in terms of where each stage can be automated:

1. *Observation stage*: the acquisition of information about particular target or a specific situational scenario;
2. *Orientation/analysis stage*: analysis of the available information;
3. *Decision stage*: making targeting decisions based upon the analysis of the available information at stage 2;
4. *Enactment stage*: enforcement of a targeting decision made at stage 3;
5. *Assessment stage*: assessment of the aftermath of the military act.

In this case, (semi-)autonomous drones typically automate the large quantities of data that their sensors input in the first two stages. Drone programming filters out what it deems irrelevant to decision-making and feeds the remainder to the pilot. The pilot may make a decision at the third stage, then feed that decision to an automated payload delivery system (stage 4). This, of course, is just an example of how various stages can be automated or not. As this paradigm can design different combinations of automation and human control, automated targeting and payload delivery is not an either/or proposition. Instead, the distinction between fully-autonomous drones and the semi-autonomous system described above is that the former ascribes automation to the entire five-stage process. Leveringhaus refers to those who program fully autonomous drones as 'drone programmers', distinguishing between them and 'drone pilots' who form the human-in-the-loop paradigm of semi-autonomous drones.

Moral Responsibility

Leveringhaus draws on the work of Santoni de Sio and Di Nucci (2016), delimiting his conception of responsibility to focus solely on *moral* responsibility rather than two other distinct (albeit interrelated) concepts of *causal* and *legal* responsibility (2016, 170). By centring moral responsibility as the focus of MHC, he adopts Strawson's (1962) conception of moral responsibility that eschews the nuanced arguments underlying debates on free will. The Strawsonian approach features notions of blameworthiness and praiseworthiness predicated on social practices (Leveringhaus, 2016, 170). The moral responsibility of any given agent ensures their liability for any praise or blame associated with the results of their practices. Other agents are similarly justified in attributing the proper praise or blame to the liable agent. To this end, Leveringhaus' (2016) chapter explores whether or not "the increasing automation of drones necessitates a rethinking of practices of praising and blaming" (170).

Just War Theory

Just War theory is used as a moral landscape to frame the practices of praising and blaming. The theory refers to the rules and regulations that constrain the use of force in any given armed theatre. Leveringhaus (2016) focuses on one of the tripartite vectors of Just War theory, *jus in bello* or justice in war (as opposed to *jus ad bellum* or justice pre-war in terms of the declaration of war, and *jus post bellum* or justice after war) (171). The three criteria for assigning responsibility for recalcitrance in *jus in bello* are as follows:

> *Distinction* obliges belligerents to distinguish between legitimate and illegitimate targets by not intentionally targeting the latter;
> *Proportionality of means* obliges belligerents not to cause excessive harm; [and]
> *Military necessity* obliges belligerents not to cause unnecessary harm. (ICRC, 1949; Leveringhaus, 2016, 171)

Leveringhaus continues by arguing that the practices of praising and blaming responsible agents provides a solid starting point for tackling these issues. However, they are insufficient for allocating full moral responsibly to the military. He further provides five conditions that must be met to identify responsible agents (thus making them liable for the blame or praise mentioned above). The first three conditions are adopted from Cowley (2014), while the fourth and fifth derive from the Nuremburg trials to note that an agent must have:

1. *Moral capacity:* the agent must comprehend what they did and why they are held responsible for such action(s);

2. *Moral understanding:* the agent must comprehend the moral context within which their actions were undertaken;
3. *Control:* the agent must have been in control of their actions (i.e., have possessed the ability to not act the way they did);
4. *Moral perception:* the agent must show they had attained (or could not have attained) the morally relevant knowledge that allowed them to assess their use of armed force in a particular context;
5. *Moral choice:* the agent could have been able to avoid executing a particular order.

(Leveringhaus, 2016, 171-172)

According to this criteria, drone pilots and programmers could ostensibly be seen as not morally responsible for their actions. This is due to the nature of the automated system, which precludes their ability to have either sufficient *moral perception* and/or *control*. However, the degree of moral competence underpinning moral perception does not preclude a programmer from understanding the basic moral rules of their domain (i.e., war theatre). Leveringhaus (2016) makes the salient point that "automated targeting does not necessarily challenge developing adequate moral competence. Whether a member of the military develops or fails to develop adequate moral competence depends, I contend, much more on training then on subsequent uses of a particular weapon" (173).

This would require sufficient training in ethics and law for members of the military, including programmers, to understand the underlying elements of *jus in bello*. Automated targeting does not exclude this moral competence *per se*, as a programmer must be aware of the principles of distinction and proportionality. When designing a system, for instance, the programmer who ignores these principles would be in direct contravention of *jus in bello*. Of course, and as Leveringhaus admits, there is a gap between agent comprehension of the relevant rules and the actual application of these rules. To illustrate, automation of step 1 (observation) and step 2 (orientation/analysis) of the kill chain is problematic. This is because the filtering of collected information and subsequent feeding-up to the pilot limits the relevant moral knowledge necessary for proper analysis. The moral perception of the pilot would thus be hindered, affecting decisions made in the remaining stages.

Yet one might also argue the opposite: without such filtering, the sheer volume of large quantities of data could equally obscure the moral perception of pilots and hinder morally relevant decision-making. This becomes a fundamental point in the design architecture for these systems. The automation of stage 3 (decision) could delimit the choices of available targets for a pilot to act upon. However, it could also delimit the space for human error to occur. Similarly, automation of stage 4 (enactment) could limit the ability of a pilot to intervene in payload delivery. But it could also carry out such a

delivery with greater precision than a human pilot could. In these semi-automated scenarios with human pilots, the landscape of automation is complex and nuanced. The moral issues that arise become even more problematic when we consider full automation of the kill chain.

Since full automation of the kill chain by an AWS precludes the moral perception of programmers, there are arguments about programmer ability to design a kill switch that could be used to intervene in recalcitrant systems (Contissa, Lagioia, & Sartor, 2017; Leveringhaus, 2016). This would put full moral responsibility back in the hands of programmers, given their ability to intervene in such an absolute way. Still, distant war theatres are often complex and there are always limitations on programmer ability to attain relevant knowledge on any given deployment scenario. This limits agent ability to have sufficient moral perception in turn, making moral control highly improbable. Leveringhaus argues that popular as it maybe be among AWS sceptics, such an argument fails to undermine the attribution of responsibility to programmers. Through proper outlining and the application of 'standards of care', it is possible for the military to both accept the assignation of responsibility and adhere to it.

It is nonetheless difficult for programmers to have sufficient moral perception of what AWS do during deployment. Leveringhaus argues that they should take a forward-looking approach to moral responsibility, assessing the potential risks that may arise from automation of the kill chain once deployed. Drawing from risk theory (which he argues is underdeveloped in Just War theory), Leveringhaus believes that riskless war is impossible. Yet programmers can nonetheless account for various possible risks that could emerge, and balance these risks against each other through design decisions. If this is the case, then programmers are still responsible for the resulting risks associated with outcomes from the use of automated targeting systems (Leveringhaus, 2016, 176).

Because the programmer is aware of their limited perception of morally relevant facts and risks during deployment, they retain moral responsibility for the decisions they make in terms of mitigating and reducing risks prior to deployment. Moral perception of an AWS during deployment is critical to understanding risk in warfare. For programmers in particular, moral awareness of the risks imposed by deployment is crucial to understanding whether their imposition of the associated risks is *justified, negligent,* or *reckless* (Leveringhaus, 2016, 176). If it can be shown that the imposition of risk was justified, then associated actions and outcomes do not merit blame (they also do not necessarily merit praise). If the associated risks of deployment were negligent or reckless, then it could be said that the moral perception and/or competence of the programmer(s) was lacking. If it can be shown that such agents failed to take sufficient steps to acquire morally relevant knowledge before making decisions

about risks, then they could be held responsible for wrongdoing. Thus, their actions would merit blame. This forward-looking approach to moral responsibility (which is also a fundamental precept of VSD, discussed in greater detail in Annex II) supports a broader moral perception. Leveringhaus (2016, 177) argues that this approach actually aligns with the equally broad notion of the Standards of Care (SoC).

The SoC approach is predicated on devising and adhering to sound principles of care regarding the responsible use of semi- and fully autonomous targeting systems in AWS. It is intended to determine the contexts for use wherein the deployment of such systems impose reasonable risks. Resistance to automation as a danger *per se* is eschewed here, as responsible use of automation is contingent on relevant contexts for deployment and the standards of care used in such contexts. This means the SoC approach provides a landscape within which moral responsibility can be assigned to programmers and pilots. Failure to adhere to either an existing standard of care or a sufficiently adequate standard of care can provide the basis for blameworthiness. In other words,

> [s]tandards of care would also govern interactions between drone pilots and drone programmers. To reduce risk, drone programmers would have to be *transparent* about the ways in which they program partially automated drones. They would have to inform their colleagues about the parameters being used for automation, as well as the stages of the targeting process being automated. (Leveringhaus, 2016, 177, emphasis mine)

Within the military context in particular, standards of care also apply to the superiors of pilots and programmers. It is the duty of these superiors, along with the more general military apparatus, to assess the efficacy of existing standards of care. It is also their duty to develop and implement sufficient standards for these types of automated technologies – and arguably for all emerging technologies.

To sum up, Leveringhaus recognises that the deployment and continued development of automated technologies (such as the AWS described above) is not unproblematic. But problems arising from development can nonetheless be addressed through revision of what constitutes moral blameworthiness and praiseworthiness. MHC is then achieved by adhering to standards of care sufficient to reduce negligent and reckless risk-taking, as per the conditions set out above by relevant moral actors. Actors include not only pilots and programmers, but also their superiors and embodying institutions.

2.4 MHC as Design Requirements[5]

[5] Much of description in this section is adapted from a paper I previously published, which similarly recounts Santoni di Sio et alia account of MHC (Umbrello, 2020).

In their seminal 2018 paper titled "Meaningful human control over autonomous systems: a philosophical account," Santoni de Sio and van den Hoven depart from existing accounts for MHC to instead provide a philosophical one. Their account defines MHC as co-variance between system behaviour and an agent's decisional intentions and reason to act. The approach aligns directly with (and emerges from) responsible design practices and value sensitive design (VSD), above all (Santoni de Sio & van den Hoven, 2018). This means systems can be designed in ways that permit agents to forfeit some of their direct operational control while still retaining global control over the system. Ironically, more – not less – levels of autonomy may permit greater control over a system in some cases. Santoni de Sio and van den Hoven (2018) provide the salient and timely example of autonomous vehicles or self-driving cars, where users retain overall control of the autonomous mobility system even though the system can conceivably put the user in unforeseen and potentially threatening conditions. Attaining MHC in this sense allows for clearer lines of accountability to be drawn when humans remain 'in-the-loop' over a system, as tracking the relevant reasons behind agent decisions is a necessary condition.

This approach to tackling MHC is novel because it is comprehensive in its scope, looking beyond discrete systems to the entire sociotechnical infrastructure to which these systems belong. Although the specific design and deployment of a system implicates important factors for understanding MHC, it cannot be understood in isolation from the infrastructure, organisations, and other agents that are inextricably connected to system design, deployment, and use. The approach is also novel because it frames MHC as capable of being designed by engineers – that is, as technical design requirements not only for the system itself, but also for the larger sociotechnical infrastructure. But in order to achieve this, two conditions must be met: *tracking* and *tracing*. As we shall see below, satisfaction of these two conditions allows for a more expansive, comprehensive notion of meaningful human control. This notion extends beyond solely users to permit agents (such as designers, policymakers, organisations, and states) to exert a level of meaningful control. It thus demarcates clearer lines for the attribution of responsibility.

Tracking and Tracing Conditions

Building off Fischer and Ravizza's (2000) concept of reason-responsiveness in their theory of moral responsibility, Santoni de Sio and van den Hoven (2018) propose two necessary conditions for

MHC: *tracking* and *tracing*. The tracking condition deals with how responsive a system is to the actions resulting from human rationale.[6] It is more comprehensively defined as the

> *[f]irst necessary condition of meaningful human control*. In order to be under meaningful human control, a decision-making system should demonstrably and verifiably be *responsive* to the *human* moral reasons relevant in the circumstances – no matter how many system levels, models, software, or devices of whatever nature separate a human being from the ultimate effects in the world, some of which may be lethal. That is, decision-making systems should *track* (relevant) human moral reasons. (Santoni de Sio & van den Hoven, 2018, 7)

In order for a (semi-)autonomous system to satisfy the tracking condition, its behaviour must map onto the reasons (intentions, plans, objectives, etc.) causing the relevant human agent(s) to undertake or abstain from any action. The tracking condition, then, is contingent to determinant design requirements. It requires an autonomous system such as an autonomous vehicle to be designed so that, after taking into account all accessible relevant input, system behaviour corresponds with human reasons for (in)action as much as technically possible. If system behaviour co-varies coherently with the (moral) reasoning of an agent, then the system can be said to fall under MHC.

The tracing condition differs in that it examines whether it is possible to determine the human agent(s) within the history of system design and deployment (e.g., designers, manufacturers, users, etc.) who are capable of understanding the system's potential and recognising their moral responsibility for the use or deployment of the system (i.e., the liability of moral consequence). Santoni de Sio and van den Hoven (2018) define tracing more thoroughly as the

> *[s]econd necessary condition of meaningful human control*: in order for a system to be under meaningful human control, its actions/states should be traceable to a proper moral understanding on the part of one or more relevant human persons who design or interact with the system, meaning that there is at least one human agent in the design history or use context involved in designing, programming, operating and deploying the autonomous system who (a) understands or is in the position to understand the capabilities of the system and the possible effects in the world of the its use; (b) understands or is in the position to understand that others may have legitimate moral reactions toward them because of how the system affects the world and the role they occupy. (9)

MHC is attained by agents who can satisfy both of these conditions; only then can they be said to have MHC over a system. AWS can *prima facie* fall under MHC through one or more agents when they are designed to support the values of accessibility and explicability (explainability and transparency). These values should manifest in system behaviour as much as possible. If a system is able to explain its

[6] The use of the term 'reasons' here is understood as any element that can both prompt and demonstrate human behavior, such as objectives, programs and strategies.

internal decision-making (explicability) and such systems are themselves transparent (also a factor of explicability), then such systems can – at least in theory – be brought under MHC more easily. This is because agent understanding of system use and deployment can be more easily attributed to the architecture of system design.

With these two necessary conditions, MHC ultimately entails a definition of control that is both more nuanced and more stringent than operational control, which demands full direct control. It is more stringent than direct control in that it precludes the attribution of human control to systems just because they have an agent 'in-the-loop' (e.g., a soldier co-commanding a field operation with a AWS). Even if armed with a kill switch and visibility of the current status and activities of the AWS, a commander of a AWS is not necessarily equipped to understand why the system does what it does. Many autonomous technologies are subject to 'black boxing', which is when the technical infrastructure of a system makes its inner workings opaque to the user. In such cases, MHC by the end user cannot be attained because the tracing condition cannot be met due to system opacity. It is true that other agents, such as designers, programmers, the military institution, or even the state, may very well understand what is going on in the so-called black box (although not always). Responsibility or MHC can be attributed to these agents as follows: if the system successfully tracks their reasons, and if agents are responsible for and capable of understanding the behaviour that the system exhibits (based on that tracking), and if agents are also responsible for the way it acts (based on its tracking of more proximal reasons discussed below).

This understanding of MHC is more comprehensive than that of direct operational control because it permits the inclusion of supervisory control. Supervisory control sanctions the user to supervise an (semi-)autonomous system that is under operational control, yet also allows the user to intervene in operations if necessary. At the same time, this form of direct supervisory control is not a necessary condition for possessing MHC. In principle, a (fully) AWS can be precise, comprehensive, and transparent in tracking the reasons behind the decisions of a human agent in lieu of the human ability to intervene in operations. This would still meet conditions for MHC.

Distal and Proximal Reasoning

Santoni de Sio and van den Hoven further develop this conception of agent reasoning adopted from the philosophy of intent and action (Bratman, 1984; Mele & William, 1992). Their development helps in not only specifying types of reasons within complex systems, but also better understanding the inner workings of the tracking condition detailed in Calvert et al. (2018). Calvert et al. (2018) began by

identifying two types of reasons: *distal* and *proximal*. Proximal reasons are those intentions that adjoin an action in a temporally immediate (concurrent) way. For instance, an agent might intend for a system to fire on an enemy combatant in order to cover their flank or to prepare for a dynamic breach. Distal reasons are longer term intentions or objectives that are formulated in a less immediate way. The use of AWS to reduce allied casualties or to increase operational effectiveness, for example, is a distal reason.

	Distal Reasons (longer term, general objective)	**Proximal Reasons (concurrent intentions)**
(Semi-)Lethal Autonomous Weapons	• Plan to maximise operation efficiency • Plan to reduce human casualties	• Intention to fire on acquired target • Intention to move to exfiltration area

Table 2. Example of distal and proximal reasons with regards to AWS

Distal reasons are the overarching intentions that a relevant agent(s) has for desired system operations. The concept of direct operational control is naturally aligned and sensitive to proximal reasons, wherein a system functions as a consequence of the immediate, concurrent intentions of the human agent. If the pilot of a (semi-)autonomous drone does not fire on one or more acquired targets, for example, it is because the pilot had no intention to do so in that instant. Perhaps the pilot was waiting for more information from ground troops or looking for reinforcements. Semi-autonomous systems like these are, to the best extent possible, influenced by the proximal reasons of their human users (pilots). Those users are thus causally responsible for the use and consequent impacts of a system.[7] MHC expands the scope of reasons a system must be sensitive to in order to sufficiently satisfy the tracking condition. We can assume AWS are likely connected to various other autonomous systems (such as satellite tracking, non-lethal support/operational unmanned ground vehicles such as DRDO Daksh, information communication technologies, and warning systems). They must thus be sensitive to both proximal reasons as well as distal ones. Satisfaction of solely proximal reasons (such as firing on target) can sacrifice more general, objective, and distal reasons (such as the reduction of civilian causalities).

[7] This is debatable, given the types of information fed upwards to the user through target acquisition and filtering systems programmed by designers (see Leveringhaus 2016 in the preceding section).

Part I of the dissertation discusses this 'systems thinking' approach in greater detail. The tracking condition, in particular, requires all elements part of any given system(s) to be maximally sensitive/responsive to the relevant (moral) reasons of agents whether users or otherwise. This means agents are not the only ones who bear the burden of demonstrating maximal ability to behave according to patterns of reasoning. Instead, every point of a system's infrastructure must be similarly sensitive. This responsiveness can be framed by designers choosing the proper 'level of abstraction' (Floridi, 2017) in creating autonomous systems based on the context for use to ensure receiver-contextualised explanations and transparent purposes (Floridi, Cowls, King, & Taddeo, 2020). An AWS, for example, cannot simply respond to user rationale only. It must also conform to legal and social norms, such as international humanitarian and human rights law or the laws of armed conflict. Mecacci and Santoni de Sio (2019) explicitly argue that, although the tracking condition requires the system to respond to human reasoning and not to other vectors in the system, social and legal norms reflect the intentions and reasons of supraindividual agents (e.g., organisations, companies, and states) (Mecacci & de Sio, 2019, 4).

The implications of their approach are not insignificant as they appear to run contra the intuition that greater autonomy entails less MHC. AWS themselves are composed of systems (e.g., for targeting acquisition, payload delivery, information communication technology, vehicle platforms, and so on). These are then integrated to form new systems (e.g., battalions, corps, the army, the military, etc.). The task of integration requires a comprehensive and ubiquitous design that permits all systems to be maximally sensitive. Sensitivity goes beyond end user intentions and reasons for action to include societal norms as well as legal statutes and policy. As already stipulated, this means having a more stringent notion of what constitutes MHC. But a more stringent notion permits increased levels of autonomy through increased control over the system by means of design decisions and regulatory infrastructure. MHC can be achieved if systems are maximally responsive to the intentions of agents beyond end users, such as the designers, companies, and states in general.

2.5 Distributed Moral Responsibility

Like Santoni de Sio et al., Ekelhof (2019) touches on the role of designers. Like Leveringhaus, he also focuses on technical targeting procedure. But Ekelhof (2019) frames MHC as a function of military operations practice that both supports and constrains targets in operational areas. Operations necessarily constrain the 'autonomy' of systems such as AWS, just as with human soldiers. The notion of 'full' autonomy is not actually full in the sense that is often implied in discussions on autonomous

weapons systems. Autonomy is always restricted by various operational decisions and planning *a priori* to deployment and operations.

Ekelhof begins by using a case of conventional air operations to frame human operational involvement in a dynamic targeting process. The case illuminates the role of human agent decision-making within distributed systems, providing steps for decision-making about military planning and operational function. Outlining practices that contextualise the use of AWS, these steps are helpful for both policymakers and theorists. Characterising the human role in military decision-making, Ekelhof iterates a six part (pre-operational) briefing package followed by a six step landscape for mission execution. I briefly summarise these parts and steps below.

Pre-Mission

The Briefing

At this point, the air component is given mission execution information. Such information is oftentimes highly detailed in terms of "target location, times, and munitions", but also less detailed when we consider dynamic targeting *in situ* (Ekelhof 2019, 345). Information is distributed to specialists in various areas for operations who then engage in more detailed planning. The executers of the mission, in this case fighter pilots, are then brought in. Pilots are briefed on mission details and given time to study the information provided or check on any last-minute preparations. Ekelhof (2019, 345) outlines the following six components, all of which should be included in the briefing package:

1. A description of the target (such as a military compound) consisting of all available knowledge;
2. Target coordinates;
3. A collateral damage estimation (CDE) to provide the operator with an idea (not certainty) of anticipated collateral damage (NATO, 2016). In this case, the risk of collateral damage is low as long as predetermined mitigating techniques are applied;
4. Recommendations for the quantity, type, and mix of lethal and nonlethal weapons needed to achieve desired effects (i.e., a weaponeering solution; see USAF, 2017). Our example requires GPS-guided munitions;
5. The joint desired impact, which is used as a standard to identify aim points; and
6. The weather forecast. In our case, it will be an overcast night (clouds covering most or all of the sky) and heavy rainfall.

Coupled with other information such as the rules of engagement, the operator can then depart and execute the mission.

***In Situ* Operations**

Step 1: Find

Intelligence and data are required to locate the target of operations. In this case, the target is pre-programmed into the navigation systems of both the fighter jet and the payload. Whereas a dynamic target requires *in situ* data collection, here the task involves arriving at the pre-programmed "weapons envelope (i.e., the area within which the weapon is capable of effectively reaching the target)" (Ekelhof 2019, 345). This process is displayed on the operations heads up display (HUD).

Step 2: Fix

Once the operator arrives within the weapons envelope, onboard systems aim to positively identify the target that was confirmed during operational planning. This ensures payload delivery complies with relevant military and legal norms (e.g., NATO, 2016). In this case, targets were pre-planned and confirmed so the operator typically does not engage in visual confirmation of positive target identification. Instead, the operator relies on onboard systems and the validation that took place during operational planning to ensure the identified target is lawfully engaged. Even in this fixed case of pre-planning, the human pilot is not required to attend to anything else during this phase other than arrival within the weapons envelope (Ekelhof 2019, 345-346).

Step 3: Track

The operator tracks the target within the weapons envelope to ensure continuity of positive identification and provide concurrent updates as to the position/status of the target. In the event of a static target (a military compound, in this case), tracking is relatively straightforward and involves simply entering into the weapons envelope (Ekelhof 2019, 346).

Step 4: Target

During this phase, the rules of engagement, laws of armed conflict, and other relevant bodies of regulation are invoked to ensure lawful targeting and deployment. Reference to rules also ensures other considerations, such as for issues related to collateral damage or the risk factors incurred by forces. Once again, in this predetermined and validated target case, the legal and military experts who vetted the target permit the pilot to simply input relevant data into the vehicle and weapons payload delivery systems to ensure proper execution. Given the visual impairment of weather conditions in this case,

further collateral damage estimates cannot be attained as no visual confirmation is possible. Because pre-mission planning determined low estimates for collateral damage, and because that planning was conducted according to governing norms, the human pilot need not actively participate or intervene beyond piloting the vehicle into the weapons envelope (Ekelhof 2019, 346).

Step 5: Engage

Once the operator enters the designated weapons envelope, the onboard computer suggests the most opportune time to release the payload for maximum effectiveness (based on computer knowledge of the capabilities of the equipped weapons systems). Since the payload system is guided by GPS, there is no need for any other forms of targeting based on visual identification. Once weapon release is authorised by the pilot, the munitions guide themselves to the target (Ekelhof 2019, 346).

Step 6: Assess

At this point, the task becomes assessing the damage that resulted from the previous stage and determining the effects of the strike. A pilot's visual assessment can be impaired by many different factors (weather conditions, in this case). Visual assessments of collateral damage from the vantage point of a pilot may likewise fail to accurately reflect the efficacy of the strike and its consequences. For aerial engagements, ground support forces may be needed to assess the engagement more accurately (Ekelhof 2019, 346).

When considering MHC, then, it appears most of the work underlying each step falls outside the control of the pilot. This is representative of contemporary aerial operations in general. Although the pilot is seen as in direct operational control for some of the operation, such as piloting the craft to the weapons envelope and engaging in weapons release, this type of control is not meaningful in a sufficient sense. Here, the pilot arguably lacks 'cognitive clarity and awareness' of the situation they are engaging with (Article 36, 2015). This begs the underlying question of whether the pilot *actually* possesses levels of clarity and awareness that might be deemed sufficient or substantial in a meaningful way. It could be argued that the operator possesses MHC only because they were briefed pre-mission and knew the details of the operations (such as the target, the weapon's payload, and the estimated damage). To that end, various actors within the hierarchy must share some level of trust in the lawful validation of briefing details and targets as well as in the normative compliance of their engagement.

These discussions at the pilot level can provide some future insight both for operations using AWS and contemporary aerial vehicles. But Ekelhof argues that such discussions focus on the wrong subject (i.e., the operator). Instead, they should focus on how the military can possess MHC over

targeting operations as an organization. He believes that current international discussions related to AWS focus overly much on the deployment stage of AWS and their relations to operators, thus positioning the nexus for MHC between those two agents. In doing so, discussions overlook the larger covariance of the division of labour between agents within the military body that forms the decision-making process. The steps outlined above, particularly the pre-mission briefing stage with its collateral damage and proportionality assessments, are largely ignored (as echoed by Roorda, 2015 in 2.1 above). Ekelhof concludes that a distributed notion of MHC is necessary to more accurately account for the various decisions and procedures that different agents engage in prior to deployment as part of a larger process.

For this reason, different agents have different levels of control over any given vector in the process. Any sufficient conception of MHC must reflect this variation in both agent and control. Such a concept would not negate the role played by human operators, of course. Rather, it would position human operators as part of a larger distributed network for decision-making. Here, 'full autonomy' is not full in the sense that is commonly intuited; it is necessarily constrained by the larger apparatus within which it forms a part. This observation reflects the point made by Santoni di Sio et al., which is that tracking alone does not necessarily entail MHC. MHC must be located post-deployment with the end user, but also with designers and CEOs as well as supraindividuals such as companies, organisations, and states (i.e., the military).[8] This echoes (and Ekelhof repeats it as well) the Defence Science Board's statement that "there are no fully autonomous systems just as there are no fully autonomous soldiers, sailors, airmen or Marines" (USSB, 2012, 23).

3 Conclusions

This introduction provides a short overview of the growing literature on meaningful human control, particularly as pertains to AWS. I have presented five different conceptual framings for MHC appearing within the literature. Each framing shares some similarities, but all are markedly different in the loci of their focus on how MHC might be achieved in these types of systems. Roorda (2015) looks at how MHC can be achieved through closer examination of the flow of targeting procedures within these systems (2.1). Ekelhof (2019) takes a more meta-level approach, scrutinising the overall targeting and decision-making apparatus of the military body (2.5). Asaro (2012) employs a more technical line of inquiry, positioning MHC as a function of agent ability to intervene in near-term decision-making

[8] The work of Santoni di Sio et al. may provide ways to design Ekelhof's conception of human-machine relationships regarding LAWS.

(2.2). Saxon (2016) aims at a similarly agent-centric approach by proposing proper commander training on the use and capabilities of these types of systems to close the responsibility gap. Leveringhaus (2016) and Santoni de Sio and van den Hoven (2018) take markedly different approaches to looking at the moral responsibility of designers as collaborators within the military sphere. Both focus on how they design these systems. They also look at tracking and tracing the moral reasons of relevant agents within the design and use of these systems. Although there may be a tendency to look at these different approaches as mutually exclusive, their differences highlight areas that actually bolster each approach. As this dissertation aims to create a more holistic conception of MHC that can be adopted to confront imminent issues with these emerging systems, the sections that follow will adapt many of the technical and philosophical underpinnings of these approaches.

References

Adams, T. K. (2001). Future warfare and the decline of human decisionmaking. *Parameters*, *31*(4), 57–71.

Allen, C., & Wallach, W. (2014). Moral Machines: Contradiction in Terms or Abdication of Human Responsibility? In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 55–68). MIT Press.

Altmann, J. (2005). *Military Nanotechnology: Potential Applications and Preventive Arms Control*. *Military Nanotechnology Potential Applications and Preventive Arms Control*. Oxon: Routledge. https://doi.org/10.4324/9780203963791

Article 36. (2015). Killing by machine: Key issues for understanding meaningful human control. Retrieved January 28, 2020, from http://www.article36.org/weapons/autonomous-weapons/killing-by-machine-key-issues-for-understanding-meaningful-human-control/

Asaro, P. (2009). Modeling the moral user. *IEEE Technology and Society Magazine*, *28*(1), 20–24. https://doi.org/10.1109/MTS.2009.931863

Asaro, P. (2012). On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross*, *94*(886), 687–709.

Asaro, P. (2016). Jus nascendi, robotic weapons and the Martens Clause. In *Robot Law*. Edward Elgar Publishing.

Bratman, M. (1984). Two faces of intention. *The Philosophical Review*, *93*(3), 375–405.

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

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

Cowley, C. (2014). *Moral responsibility*. Acumen Publishing. Retrieved from https://www.cambridge.org/core/books/moral-responsibility/CAA4DE7D46EBF47A9E43835A2BA00F64

Crootof, R. (2016). A Meaningful Floor for Meaningful Human Control. *Temp. Int'l & Comp. LJ*, *30*,

53.
Docherty, B. (2012). *Losing humanity: The case against killer robots*. Human Rights Watch. Retrieved from https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots
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
Floridi, L. (2017). The logic of design as a conceptual logic of information. *Minds and Machines*, *27*(3), 495–519.
Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). Designing AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*, 1–26. https://doi.org/10.1007/s11948-020-00213-5
Geiss, R. (2015). *The international-law dimension of autonomous weapons systems*. Friedrich-Ebert-Stiftung, International Policy Analysis. Retrieved from http://eprints.gla.ac.uk/117554/
Germany/France. (2017). *CCW/GGE.1/2017/WP.4 Working Paper for consideration by the Group of Governmental Experts on Lethal Autonomous Weapons Systems*.
Germany. (2014). *General Statement by Germany, CCW Expert Meeting Lethal Autonomous Weapons Systems*. Geneva. Retrieved from https://www.unog.ch/80256EDD006B8954/(httpAssets)/97636DEC6F1CBF56C1257E26005FE337/$file/2015_LAWS_MX_Germany.pdf
Heyns, C. Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, Pub. L. No. A/HRC/23/47, Human Rights Council (2013). United Nations General Assembly. Retrieved from http://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf
ICRC. Geneva Convention Relative to the Protection of Civilian Persons in Time of War (Fourth Geneva Convention) (1949). Retrieved from https://www.refworld.org/docid/3ae6b36d2.html
ICRC. (2016). *Expert Meeting: Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*.
ICRC. (2018). Treaties, States parties and Commentaries: General Protection of Civilian Objects. Retrieved March 24, 2020, from https://ihl-databases.icrc.org/ihl/WebART/470-750067
Kania, E. B. (2017). Battlefield Singularity. *Artificial Intelligence, Military Revolution, and China's Future Military Power, CNAS*.
Korpela, C. (2017). Report of the 2017 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS).
Leveringhaus, A. (2016). Drones, automated targeting, and moral responsibility. *Drones and Responsibility: Legal, Philosophical, and Socio-Technical Perspectives on the Use of Remotely Controlled Weapons*, 169–181.
Marauhn, T. (2018). Meaningful Human Control–and the Politics of International Law. In *Dehumanization of Warfare* (pp. 207–218). Springer.
Mecacci, G., & de Sio, F. S. (2019). Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics and Information Technology*, 1–13.
Mele, A. R., & William, H. (1992). *Springs of action: Understanding intentional behavior*. Oxford University Press on Demand.
Morley, J. (2015). Meaningful Human Control in Weapons Systems: A Primer. *Arms Control Today*, *45*(4), 7.
NATO. (2016). *NATO Standard AJP-3.9 Allied Joint Doctrine for Joint Targeting*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/628215/20160505-nato_targeting_ajp_3_9.pdf
O'Connell, M. E. (2013). Banning autonomous killing.
Roorda, M. (2015). NATO's Targeting Process: Ensuring Human Control Over and Lawful Use of

‘Autonomous’ Weapons. *Mark Roorda, NATO’s Targeting Process: Ensuring Human Control Over (and Lawful Use of)‘Autonomous’ Weapons, in: Autonomous Systems: Issues for Defence Policymakers, Eds. Andrew Williams and Paul Scharre, NATO Headquarters Supreme Allied Command Transforma*.

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

Saxon, D. (2016). Autonomous drones and individual criminal responsibility. *Drones and Responsibility: Legal, Philosophical, and Socio-Technical Perspectives on the Use of Remotely Controlled Weapons*, 17–46.

Scherer, M. U. (2015). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harv. JL & Tech.*, *29*, 353.

Stephanidis, C., Salvendy, G., Antona, M., Chen, J. Y. C., Dong, J., Duffy, V. G., … Fu, L. P. (2019). Seven HCI grand challenges. *International Journal of Human–Computer Interaction*, *35*(14), 1229–1269.

The Campaign To Stop Killer Robots. (n.d.). Retrieved March 21, 2020, from https://www.stopkillerrobots.org/

Tucker, P. (2017). Russia to the United Nations: Don’t Try to Stop Us from Building Killer Robots. *Defense One*, *21*.

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

USAF. (2017). *Annex 3-60 Targeting*. Retrieved from https://www.doctrine.af.mil/Doctrine-Annexes/Annex-3-60-Targeting/

USSB. (2012). *Defense Science Board Task Force Report: The Role of Autonomy in DoD Systems*. Washington, DC. https://doi.org/ADA566864

Walsh, J. I. (2015). Political accountability and autonomous weapons. *Research & Politics*, *2*(4), 2053168015606749.

Annex II: Value Sensitive Design and Responsible Innovation – A Literature Review

1 Methods

As mentioned in the introduction, this literature review is written according to a more comprehensive guide on dissertation writing that Randolph (2009) explores in *Practical Assessment, Research and Evaluation*. Although I considered other strategies for conducting the literature review, I noticed Randolph's guide captured overlapping tools and strategies. I thus decided to adopt this more comprehensive approach. The guide sets out a strategy for producing a satisfactory literature review that normally takes six months. Given the time allowance for the completion of a doctoral dissertation (as opposed to some other project), this approach was carried out in full. Although there are some limitations to this review, they are outlined in the inclusion/exclusion criteria described below.

1.1 Keywords

An iterative abductive process was necessary to identify relevant keywords. The process began with creating a *prima facie* list of potentially relevant keywords. Next, sources using those keywords were identified iteratively. Keywords were then modified based on the relevance of these sources. Sources that were too specific were reviewed in-depth for relevance, while overly general sources were reviewed based on sources that cited them. This process was adopted in light of the history and overall volume of literature that would have emerged if the set of keywords had been too large. To ensure the

quality and relevance of selected literature, I chose three keywords that were used either independently or together in some combination

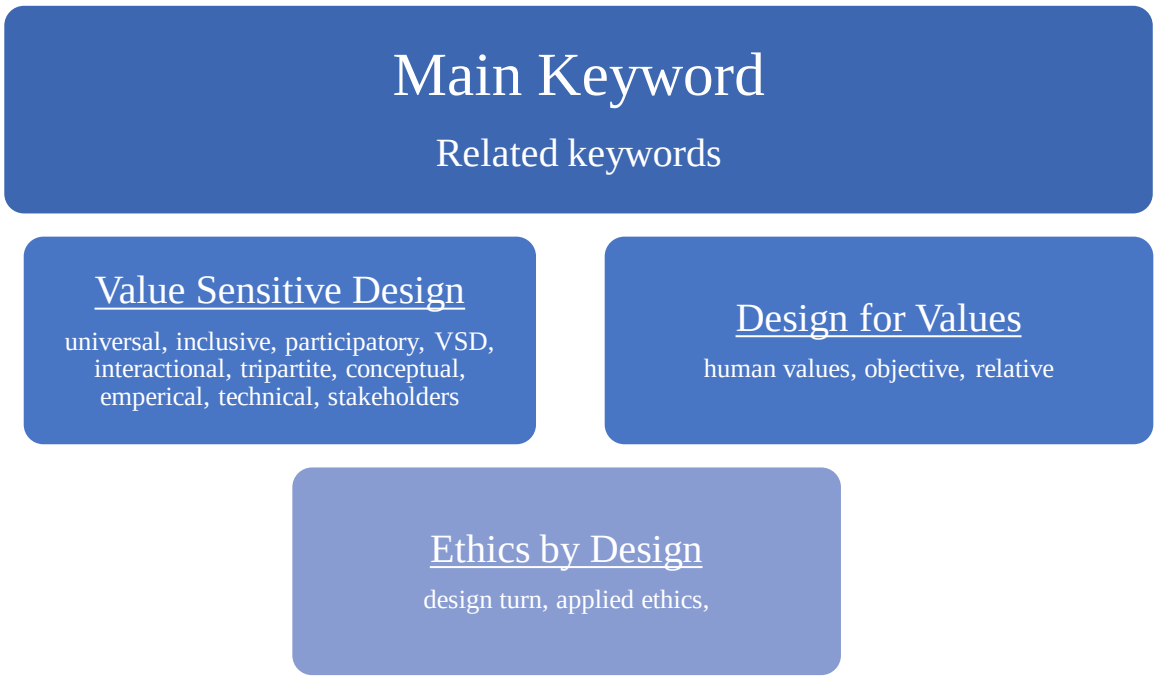

Figure 1. The three main keywords listed with more expansive keywords underneath. Some sub-keywords are more specific, while others are more general. This is because certain keywords are transdisciplinary yet often carry connotations.

1.2 Research Coverage

The coverage of this literature is an "exhaustive review with selective citation" (in Cooper, 1988 as cited by Randolph, 2009). The aim was to formulate a comprehensive list of scholarly articles relevant to 'value sensitive design'. Due to more than two decades of research into the topic, there is a multitude of sources. Many sources discussed theoretical conceptions of VSD, and a smaller population of those articles addressed the application of VSD. Although this review mentions many of these articles, most are not discussed in depth due to the specificity of their content. Instead, the review extracts the literature on VSD as a whole to give a broader and more comprehensible understanding of the approach.

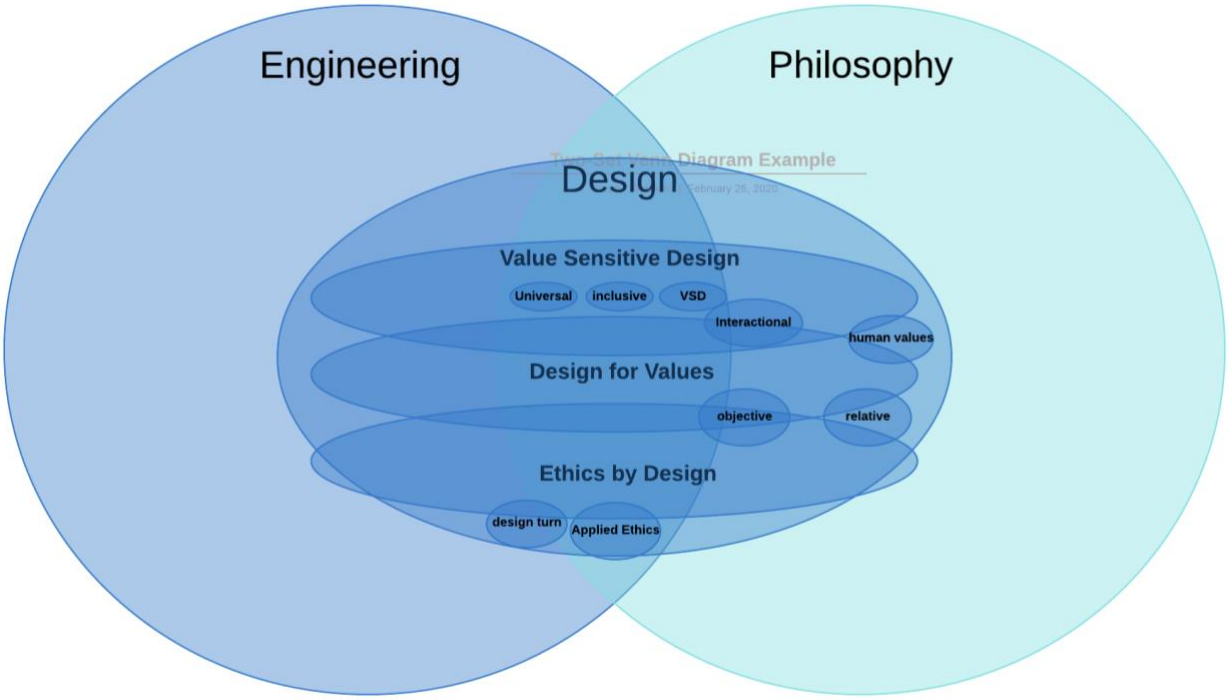

Figure 2. An innate tension exists within the keywords relevant to the literature chosen for review. Given the (trans-)interdisciplinarity of potentially relevant literature, where should a review begin? The keywords 'Value sensitive design', 'design for values', and 'ethics by design' are shown in their respective groupings. Groupings are located in their primary umbrella divisions of 'engineering' and 'philosophy'. The terms 'universal', 'inclusive', 'VSD', and 'interactional' fall primarily within the cross domain of 'design'. The terms 'human values', 'objective', and 'relative' fall primarily within 'philosophy'. The terms 'design turn' and 'applied ethics' then fall within the intersection of 'engineering' and 'design'. The intersection of 'Value sensitive design' falls within the area of overlap between the three main branches: the two umbrellas of 'engineering' and 'philosophy' and their nexus point of 'design'. Given its centrality, this term proved most salient and relevant for identifying sources of literature to include here.

1.3 Research Focus

Cooper (1988) outlines four different research focuses that can be emphasised as the foundation for literature taxonomy: research methods, theories, application, and research outcomes. The focus chosen here is theories, as the literature that forms this group provides overviews and theoretical explorations on what constitutes VSD. Centring this particular category of literature will help the reader better understand the theoretical commitments and tools underpinning the VSD approach. It can also highlight some conceptual weaknesses, which are discussed in Part II of this dissertation. This does not mean the review excludes all other categories of VSD literature. On the contrary, Randolph (2009) argues that all literature reviews are some amalgamation of different categories. This review draws on literature from the other three categories as case studies and examples to further enrich the theoretical unweaving of VSD methodology.

1.4 Inclusions and Exclusion Criteria

Given more than two decades of rich inquiry into the topic, any single review of the literature must narrow down the scope of inclusion to sources that best convey both the history and state of the art. It must also selectively exclude sources that may be redundant or less-than-relevant. The following list of criteria for inclusion/exclusion is informed by Randolph (2009):

1. Only English sources;
2. Only publications in academic journals and books;
3. Only sources from PhilPapers, the Association for Computing Machinery (ACM), the Institute of Electrical and Electronics Engineers (IEEE), Springer databases (excluding patents and citations) and vsdesign.org[9];
4. Only sources that included 'value sensitive design', 'value-sensitive design' or 'vsd' in the title, abstract, or as a keyword.[10]

The stringent nature of these four criteria ensured a more tractable body of literature overall. References to only academic journal articles and books further ensured that the source material was of a higher quality.

1.5 Overview of Final Sources

The review looks at literature from 1996 (for publication of the earliest paper on VSD, see Friedman, 1996) and December 2020. I identified 417 article and book contributions in total, of which only 57 were available. I reviewed and short-listed all 57 contributions based on their abstracts and contents. To create the basis for this review, I chose fifteen sources that best encompassed and exemplified either the theoretical commitments or practical applications of VSD. These sources are listed in Appendix I. Figure 3 below illustrates the marked rise in VSD literature spanning the search parameters.

[9] vsdesign.org is a website managed by the Value Sensitive Design Research Lab and its directors Batya Friedman and David Hendry, the pioneers of VSD.

[10] Boolean search strings using the keywords ensured fidelity in the results. Relevance weights of a value of '3' were used (the keyword must appear at least three times among the search domain parameters) to increase the probability of relevant results; see Annex II.

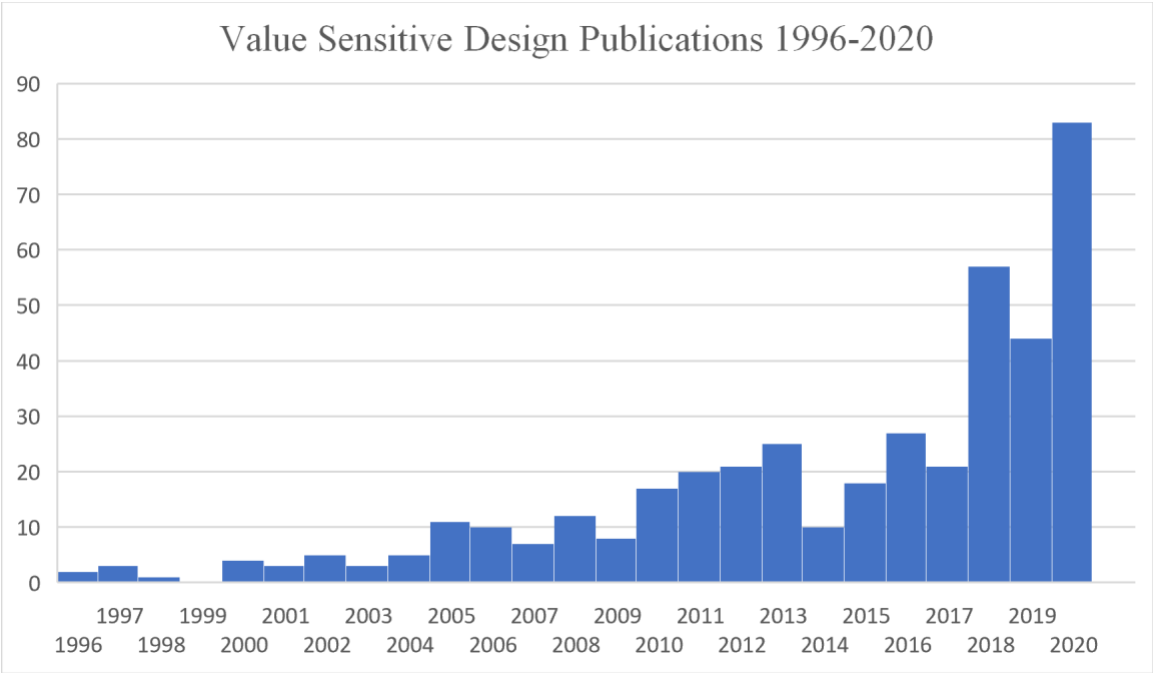

Figure 3. Value Sensitive Design Publications 1996-2020

Researchers from the University of Washington and the technical universities of Delft and Twente often appear in searches and form the bulk of Appendix I. These include Friedman (1996), van de Poel (2014), van den Hoven and Manders-Huits (2009), and Vermaas et al., (2010) among others. Their works are published primarily in edited collections by Springer or in technology journals such as *Science and Engineering Ethics, Ethics and Information Technology, Philosophy and Technology*, engineering journals such as *ACM Transactions on Computer-Human Interaction*, and trade journals from the Institute of Electrical and Electronics Engineers (IEEE).

1.6 Research Questions

As mentioned earlier in the research focus section, Cooper (1988) provides guidelines for organising and formulating research questions along the four axes of research methods, theories, practices, and research outcomes. This literature review focuses primarily on theories but incorporates literature that can be arguably assigned to the other three categories in support (or as examples of VSD in various domains). Although relevant, my own articles on this topic are excluded from this review beyond cursory mentions as the relevant ones are reproduced in full in later sections of this dissertation.

1.6.1 Theories

Q1.1: What is the origin of VSD and why was it developed?

Q1.2: How has VSD theory changed over time and what has it gained?

1.6.2 Research methods

Q2.1: What are VSD method(s) and how are they developed?

Q2.2: What makes VSD different than other stakeholder-based approaches to design?

1.6.3 Practices

Q3.1: How successful has the VSD approach been to *actual* design programs?

Q3.2: Do speculative VSD applications have any tangible boons?

1.6.4 Research outcomes

Q4.1: Do VSD applications to real-world design programs show sustainable results that might promote adoption?

2 Results and Discussion

2.1 Theories

Developed in the late 1980s by Batya Friedman and Peter Kahn at the University of Washington, VSD grew out of the field of human-computer interaction (HCI) and information systems design. It emerged as a theoretically grounded – often termed 'principled' – approach to incorporating human values into technology design through stakeholder elicitation (Friedman, 1996). There is growing adoption of a philosophical stance on human-technology interaction and relations. This stance argues that technology is neither purely deterministic nor value-neutral (instrumental). Instead, it is imbued with values. Many stakeholder theories and methods have thus emerged as a consequence of trying to design beneficial technologies (Davis and Nathan, 2014; Friedman, 1996; Friedman et al., 2013a; Manders-Huits, 2011; van den Hoven et al., 2015). Due to its more comprehensive consideration of human values, which occurs in the early stages of the design process as well as throughout, VSD has gained the most traction over the last two decades (Friedman et al., 2015, 2013a; Winkler and Spiekermann, 2018). In an abstract, Friedman and Kahn Jr. (2002, 1) best summarise VSD as "a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process. It employs an integrative and iterative tripartite methodology, consisting of conceptual, empirical, and technical investigations."

The founders of the approach devised the methodology in response to a longstanding need within the HCI community to incorporate human values (which were already of significance) into their design programs. Researchers within the HCI sphere had already considered values such as privacy (Ackerman and Cranor, 1999; Agre, 1997; Fuchs, 1999; Jancke et al., 2001; Palen and Grudin, 2003; Tang, 1997), ownership and property (Lipinski and Britz, 2000), physical welfare (Leveson, 1991), freedom from bias (Friedman and Nissenbaum, 1996), universal usability (Shneiderman, 1999; Thomas, 1997), autonomy (Suchman, 1993; Winograd, 1993), informed consent (Millett et al., 2001), and trust (Fogg and Tseng, 1999; Palen and Grudin, 2003; Rocco, 1998; Zheng et al., 2001). However, there was no overarching approach that might enable practically minded designers to design *for* human values rather than sidelining them as an afterthought for *ex post facto* and *ad hoc* additions (Friedman and Kahn Jr., 2002).

Support for VSD adoption by designers argues that the approach is fundamentally predicated in method, aiming to bridge the gap between abstract stakeholder values and more tangible design requirements that designers can operationalise through design. Descriptive accounts of methods can help designers and theorists systematically compare the boons and privations of competing approaches. As with other approaches to technological design, VSD methods emerge from underlying theory. They can be used reflexively to clarify and strengthen the theory, forming a recursively self-improving design practice.

VSD is grounded in the notion that human values do not exist in isolation from their socio-cultural contexts. Design programs themselves emerge from, or are situated in, these same socio-cultural contexts. Values such as privacy and security can be understood in different ways by different people (van den Hoven and Weckert, 2008). This position directs design teams to consider human values as part of a larger sociotechnical milieu, inextricably tied to situated human practices. It permits a more comprehensive understanding of how technical harms and benefits can be understood during design phases through sociocultural contextualisation (Friedman et al., 2017a). The following subsection discusses the six[11] main philosophical underpinnings that help to frame the VSD approach and associated methods: 1) an interactional stance on technology, 2) a tripartite methodology, 3)

[11] Friedman et al., (2017) list seven philosophical underpinnings of VSD; the one excluded here is 'Co-evolving Technology and Social Structure'. I excluded it as a discrete tenet of the approach given that it is a direct consequence of 1), the *interactional stance on technology*. This stance argues that understanding technology and society cannot be done in isolation. Instead, systems are sociotechnical. Separation of technology from society is a disservice, making our understanding ontologically weaker and less comprehensive. Holistic design takes the interactional stance to mean design should be cognizant of the fact that technology and society co-evolve with one another, and that each has an inextricable effect on the other (whether through supports and/or constraints).

stakeholder enrollment, 4) engagement with value tensions, 5) consideration for multi-lifespan design, and 6) framing design as a program for progress rather than perfection (Friedman and Hendry, 2019).

1. *Interactional stance on technology*: as mentioned already, the VSD approach rejects technological and social determinism along with pure instrumentalism. Instead, it adopts a more ontologically relational approach that argues for the co-constitution of technology and human values. This means the various entities that enable technological function, such as humans, organisations, infrastructure, and cultural groups, also design technologies. In turn, technologies influence how those entities function as well as the designs of future technologies (i.e., social entities are designed by designed technologies).
2. *Tripartite methodology:* since its inception (Friedman, 1996), VSD has been described as composed of three distinct parts or 'investigations' (in other words, it is 'tripartite'). In Figure 4 below, the *conceptual*, *empirical*, and *technical* investigations that form the tripartite methodology are understood as iterative. This means they are designed to feed back onto one another, creating a more robust sociotechnical design (Friedman et al., 2017a).

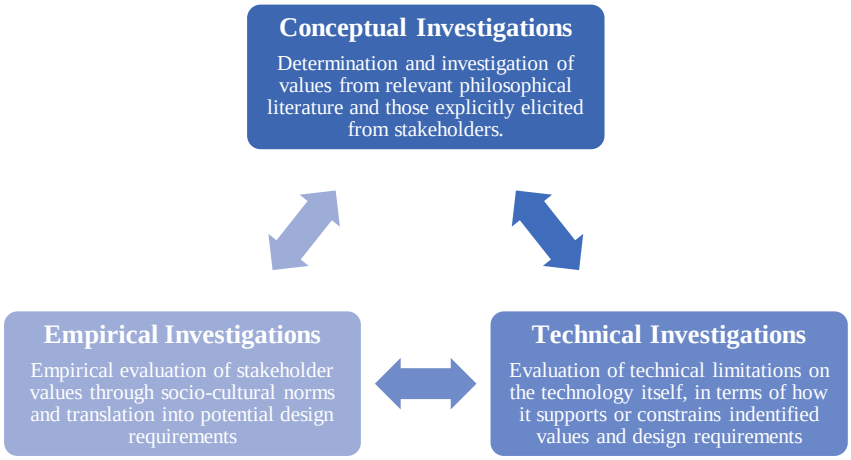

Figure 4. The recursive VSD tripartite framework employed in this study. (Source: Umbrello, 2020a)

Whether beginning with the *context of use*, the *technology*, or its *value*, the three investigations integrate philosophical, technical, and empirical approaches to stakeholder elicitation and technological design requirements. The integration bridges the gap between abstract ethical principles or values and concrete design requirements to support those values overall (see Figure 5 below). In short, conceptual investigations involve consulting available philosophical and social literature to extract some initial values and design considerations (Manders-Huits, 2011). Empirical investigations then aim to enroll

stakeholders by determining their values and design requirements. This, in turn, restructures the conceptual investigations to ensure elicited results map onto the findings (van de Poel, 2014b). Technical investigations take the technology itself under consideration in order to determine how the physical and/or digital architecture of the artifact supports and/or constrains any of those values (Friedman, 1999; Gazzaneo et al., 2020). Tensions and design requirements are formed through the integration of the three investigations. Integration occurs as each stage iteratively informs the other through the design program to create the most robust system overall (van den Hoven et al., 2012).

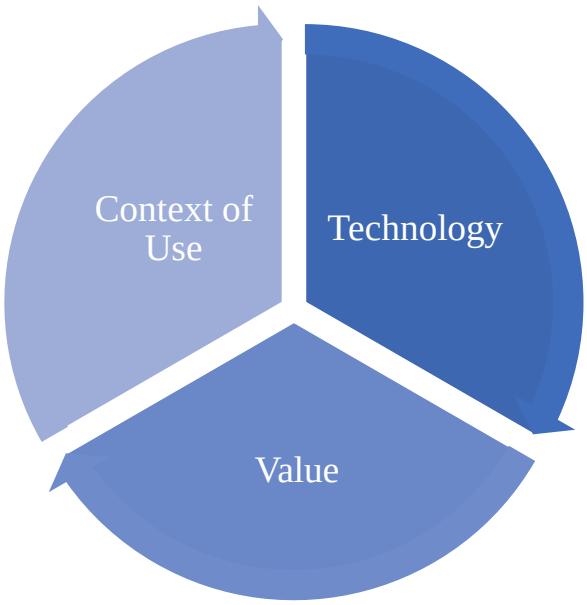

Figure 5. Starting considerations for VSD. Typically, one of the three is most pertinent to any given design. (Source: Umbrello, 2021)

3. *Stakeholders*: stakeholder enrollment is one of the most central distinguishing facets of the VSD approach. As part of the process of becoming value-sensitive, consideration of 'what' values exist in design necessitates further consideration of 'whose' values are present. Stakeholders have 'stakes' in the design and deployment of a technology. They are those who will be most affected by design decisions. Designers are obligated to enroll the most robust set of stakeholders into the design program, eliciting what values are most important to them in order to design *for* those values (Cummings, 2006). Designers must explicate the direct stakeholders (such as users) along with indirect stakeholders (such as animals, the environment, or other nations) (Friedman et al., 2013b). Because designers are making the design decisions, they must likewise be explicit about their own values as direct stakeholders. This clarifies how their decisions will either support or

constrain technical design requirements, and allows them to measure bias or take steps to de-bias their design decisions (Davis and Nathan, 2014; Umbrello, 2018).

4. *Value tensions*: as mentioned above, values are not discrete, isolated entities. Rather, there is a mesh of interconnected valuations of what stakeholders deem important. This mesh exists at any spatiotemporal point. It also emerges over time. Discussion of any given value thus implicates numerous other values and valuations, all of which must then be confronted in design. These valuations can and will exist in tension with other valuations, whether conceptually or across various scales of human experience. Individual valuations can conflict with one another, with groups, and with other populations (Friedman et al., 2017a). This is complicated even further as valuations across various scales augment over time, making the designer's task of mapping design requirements evermore difficult to nail down. Philosophically speaking, it is difficult to argue for an objective means by which these tensions might be conclusively resolved (through design or otherwise) (Davis and Nathan, 2015, 2014; Umbrello, 2020). VSD nonetheless provides a principled approach that first makes these tensions in valuations transparent, then helps guide designers through the most salient design choices around them.
5. *Multi-lifespan design*: because technology and society are inextricably linked and interact with one another, the effects of design decisions are described as 'scaffolding'. In other words, they extend into the future and (in many cases) across multiple lifespans. Cascading failures of future systems are contingent upon design decisions made in the present. Contemporary design decisions must become ever more prescient, making considerations early on and throughout design programs. Technology is implicated in the propagation of global famine, hunger, and disease. At the same time, it can be equally understood as forming part of the solution. To account for systemic effects across multiple lifespans, decisions surrounding technology design must be robust and inclusive. As Friedman et al., (2017) succinctly put it, the multi-lifespan perspective in VSD encourages "new opportunities for preserving knowledge, supporting social structures and processes, remembering and forgetting, and re-envisioning infrastructure to support inclusivity and access" (8).
6. *Progress, not perfection*: the difficulty designers face when navigating the complexities of interconnecting sociotechnical structures and designing for similarly complex human values (often in tension) makes design perfection continually elusive. In designing for human well-being, design programs must be reframed as guides towards progress. This will help provide needed solutions for pressing sociotechnical problems, and avoid the pitfalls or lag that may

come as a consequence of seeking perfection. Of course, the new framing does not exclude a *modus operandi* of continued improvement. The VSD approach simply takes this *modus operandi* as a fundamental framework for all aspects of the methodology.

2.2 Research Methods

VSD methodology encourages practitioners to engage with and operationalise the theoretical constructs underpinning the approach. It also aims for the continual improvement of investigations through practice. Similarly, the VSD approach in and of itself is not meant to be seen as a discrete or supplementary set of tools for designers to employ in any given domain. Instead, the approach is meant to be tailored and integrated into existing design environments. This intent increases its overall adoptability value (Friedman and Hendry, 2019). Over the history of VSD, scholars have proposed many methods and tools as part of the approach. In 2017, Friedman et al. proposed fourteen different VSD methods. More recently, they added an additional three methods to form a total list of seventeen (Friedman and Hendry, 2019).

In their most recent book *Value Sensitive Design: Shaping Technology with Moral Imagination*, Friedman and Hendry culminate the entirety of VSD history into a single opus. The work outlines VSD history, theory, methods, critiques, and responses along with potential future research steps (Friedman and Hendry, 2019). In the third chapter, titled 'Method', they provide a useful chart that gives the name of each method accompanied by a short overview and key illustrative references. This table is recreated below but, for the sake of brevity and to avoid redundancy, I have omitted key references to focus solely on methods.

Method	**Overview**
Stakeholder analysis *Purpose:* stakeholder identification and legitimation	Identification of individuals, groups, organisations, institutions, and societies that might reasonably be affected by the technology under investigation and in what ways. There are two overarching stakeholder categories: 1) those who interact directly with the technology or *direct stakeholders*, and 2) those indirectly affected by the technology or *indirect stakeholders*.
Stakeholder token *Purpose:* stakeholder identification and interaction	A playful and versatile toolkit for identifying stakeholders and their interactions. Stakeholder tokens facilitate identifying

	stakeholders, distinguishing core from peripheral stakeholders, surfacing excluded stakeholders, and articulating relationships among stakeholders.
Value source analysis *Purpose:* identify value sources	Distinction between explicitly supported project values, the personal values of designers, and values held by other direct and indirect stakeholders.
Co-evolve technology and social structure *Purpose*: e*xpand design space*	Expansion of the design space to include social structures integrated with technology, which may yield new solutions not possible when considering technology alone. When needed, engage with the design of both technology and social structure as part of the solution space. Social structures may include policy, law, regulations, organisational practices, social norms, and others.
Value scenario *Purpose:* values representation and elicitation	Narratives (or stories of use) intended to surface human and technical aspects of technology and context. Value scenarios emphasise implications for direct and indirect stakeholders, related key values, widespread use, indirect impacts, longer-term use, and similar systemic effects.
Value sketch *Purpose:* values representation and elicitation	A sketch of activities as a way to tap into the non-verbal understandings of stakeholders, their views, and their values as relates to a technology
Value-oriented semi-structured interview *Purpose:* values elicitation	Semi-structured interview questions as a way to tap into stakeholder understandings, views, and values as relates to a technology. Questions typically emphasise evaluative judgments (e.g., all right or not all right) about a technology along with the rationale (e.g., why?) of stakeholders. Additional considerations introduced by the stakeholder are pursued.
Scalable assessments of information dimensions *Purpose:* values elicitation	Sets of questions constructed to tease apart the impact of the pervasiveness, proximity ,and granularity of information (and other scalable dimensions). Can be used in interview or survey formats.

Value-oriented coding manual *Purpose:* values analysis	Hierarchically structured categories for coding qualitative responses to the value representation and elicitation methods. Coding categories are generated from data and a conceptualisation of the domain. Each category contains a label, definition, and typically up to three sample responses from empirical data. Can be applied to oral, written, and visual responses.
Value-oriented mock-up, prototype, or field deployment *Purpose:* values representation and elicitation	Development, analysis, and co-design of mock-ups, prototypes and field deployments to scaffold the investigation of value implications for technologies that are yet to be built or widely adopted. Mock-ups, prototypes, or field deployments emphasise implications for direct and indirect stakeholders, value tensions, and technology situated in human contexts.
Ethnographically informed inquiry on values and technology *Purpose:* values, technology, and social structure framework and analysis	A framework and approach for data collection and analysis that uncovers complex, unfolding relationships between values, technology, and social structure. Typically involves in-depth engagement in situated contexts over longer periods of time.
Model for informed consent online *Purpose:* design principles and value analysis	A model with corresponding design principles for considering informed consent in online contexts: ‘informed’ encompasses disclosure and comprehension, while ‘consent’ encompasses voluntariness, competence, and agreement. Implementations of informed consent must not pose an undue burden on stakeholders.

Value dams and flows *Purpose:* values analysis	An analytical method to reduce the solution space and resolve value tensions among design choices. Value dams are created by removing options that even a small percentage of stakeholders strongly object to from the design space. Out of the remaining design options, value flows are created from those that a good percentage of stakeholders find appealing. These are then foregrounded in the design. Analysis can be applied to the design of both technology and social structures.
Value sensitive action-reflection model *Purpose:* values representation and elicitation	A reflective process for introducing value sensitive prompts into a co-design activity. Prompts can be generated by designers or stakeholders.
Multi-lifespan timeline *Purpose:* priming longer-term and multi-generational design thinking	Primes activity for longer-term design thinking. Multi-lifespan timelines prompt individuals to situate themselves in a longer time frame relative to the present, giving attention to both societal and technological change.
Multi-lifespan co-design *Purpose:* longer-term design thinking and envisioning	The co-design of activities and processes that emphasise longer-term anticipatory futures with implications for multiple and future generations.
Envisioning Cards™ *Purpose:* versatile value sensitive design toolkit for industry and educational practice	A versatile and value-sensitive toolkit comprised of a set of 32 cards called the Envisioning Cards™. Cards are built on four criteria: stakeholders, time, values, and pervasiveness. Each card contains a title and an evocative image related to the card theme on one side. The envisioning criterion, card theme, and a focused design activity appears on the reverse. Envisioning Cards™ can be used for ideation, co-design, heuristic critique, evaluation, and other purposes.

In fact, a multitude of methodological approaches could be considered VSD. The characteristic differentiating VSD from other methodologies for technology design is that it all extends from a set of common premises. First, VSD takes an interactive stance on technology and its social-technicity nature, engaging with (direct and indirect) stakeholders through the use of a tripartite methodology in an iterative and continually improving manner. Second, VSD methods selection depends on the context of

development, the technology, or some value(s) that need(s) to be designed for. Any given method could satisfy one or more of the investigations, but methods are not mutually exclusive. Employment of more than one method may be necessary for the successful completion of the approach (Friedman et al., 2017a). Third, the various methods are not jointly exhaustive. They provide a solid starting point for undertaking a VSD approach, but may need to be substantially modified or discarded at some point in the design program as novel values, norms, constraints, and design requirements emerge in the context of innovation. To this end, the methods listed above should not be understood as static or *a priori* constrained. They should instead be seen as an outline for how to begin – one that is open to transformation. Finally, VSD should not be understood as *the* design program. It is part of a larger design practice within any given context. Hence, VSD is not hegemonic in design contexts; it is applied and adopted alongside existing design practices and norms. VSD is considered robust across various technical approaches for this reason (Friedman and Kahn Jr, 2007).

As a good starting point, VSD programs can follow eight considerations provided in Friedman et al. (2008) to put this iterative approach into practice[12]:

1. *Begin by considering a value, a technology, or the context of use.* Any one of these three core aspects easily motivates VSD. Ideally, a practitioner would begin with the one that is most explicitly and obviously critical to the designer's work or interests.
2. *Direct and indirect stakeholders.* Systematically identify direct and indirect stakeholders. Direct stakeholders are individuals who interact directly with the technology or with the technology's output. Indirect stakeholders are individuals who are also impacted by the system, though they never interact with it directly.
3. *Identify harms and benefits for each stakeholder group.* Systematically identify how each category of direct and indirect stakeholder would be positively or negatively affected by the technology under consideration.

[12] The considerations should not be construed as a concrete step-by-step method. The founders of VSD have avoided proposing step-by-step waterfall models as well as a high-level stage model for the VSD process. This was purposeful for multiple reasons. For one, they propose a revision of engineering methodologies to incorporate VSD commitments and methods by stakeholders – engineers or product managers, for example – who own or work with said methodologies. They do not believe a competing, alternative, prescriptive process will support adoption and appropriation. From the very beginning, Friedman developed VSD for appropriation by engineering and design cultures. A good example of this is the IEEE's *Ethically Aligned Design* standards report, which is part of their Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE, 2019).

4. *Map harms and benefits onto corresponding values.* Sometimes, the mapping of harms, benefits, and corresponding values will be one of identity. Other times, the mapping will be multifaceted (a single harm might implicate multiple values, such as security and autonomy).
5. *Conduct a conceptual investigation of key values.* Develop careful working definitions for each of the key values. Designers draw on philosophical literature in order to define these values more accurately and identify potential issues that already exist with certain conceptualisations of value. Investigation includes how values can be translated into norms, and how norms can then be translated into design requirements (and vice versa) (Figure 6).
6. *Identify potential value conflicts.* For the purposes of design, value conflicts should usually not be conceived of as either/or situations. Instead, they should be seen as constraints on the design space (van de Poel, 2014b). Typical value conflicts include accountability vs. privacy, trust vs. security, environmental sustainability vs. economic development, privacy vs. security, and hierarchical control vs. democratisation, among others (van den Hoven et al., 2012).
7. *Technical investigation heuristic and value conflicts.* Technical mechanisms will often adjudicate multiple (if not conflicting) values, often in the form of design trade-offs. Designers should thus aim to make explicit how a design trade-off maps onto a value conflict, as well as the differences in how it affects various groups of stakeholders (Umbrello, 2018).
8. *Technical investigation heuristic and unanticipated consequences and value conflicts.* To increase agility in responding to unanticipated consequences and value conflicts, flexibility should be designed into the underlying technical architecture in support of post-deployment modifications where possible.

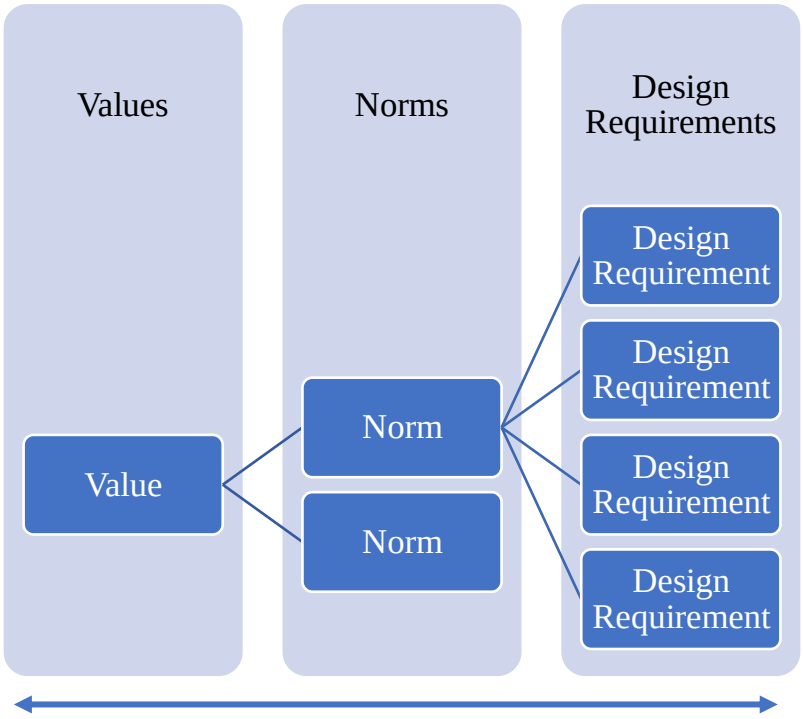

Figure 6. Bi-directional values hierarchy (Source: Umbrello, 2019)

2.3 Practices

VSD has always focused on the theoretical constructs and speculative applications of the methodologies along with their *actual* application to contemporary technologies. For example, Friedman et al. (2002) and Millett et al. (2001) describe the application of the tripartite VSD methodology in reference to a model for informed consent online. The model enrolls stakeholders in order to determine definitions and design requirements for engineering web cookies into the Mozilla browser.

Acknowledging a lack of consensus around practices to elicit user consent in terms of web browser cookies, Friedman et al. (2002) began with the value of *informed consent*. Through conceptual investigations, they went on to refine its definition as comprised of five conceptual components: disclosure, comprehension, voluntariness, competence, and agreement. With these conceptual components, they investigated how stakeholders interact with web browsers and the role that cookies play in practices. They concluded with eight design principles to help both users and designers achieve fidelity in their use and design of web browsers. Fidelity is defined in terms of attaining proper informed consent.

The eight principles are summarily listed as follows:

1. Decide whether the capability is exempt from informed consent;
2. Take particular care when invoking the sanction of implicit consent for web-based interactions.;
3. Note that defaults matter (i.e., the defaults of how a system comes to a user);
4. Put users in control of the "nuisance factor" (allow users to micro-manage their consent controls);
5. Avoid technical jargon;
6. Provide users with choices in terms of potential effects rather than in terms of technical mechanisms;
7. Conduct field tests to help ensure adequate comprehension and opportunities for agreement. (e.g., Value-oriented mock-up, prototype, or field deployment); and
8. Design proactively *for* informed consent.

A similar co-design approach to phone safety was undertaken by Yoo et al., (2013) through designer and stakeholder prompts using Envisioning Cards™ (see Figure 7). This approach allows for collaborative co-design between police offers, homeless youth, and services providers to produce more conspicuous design with user safety as an underlying value.

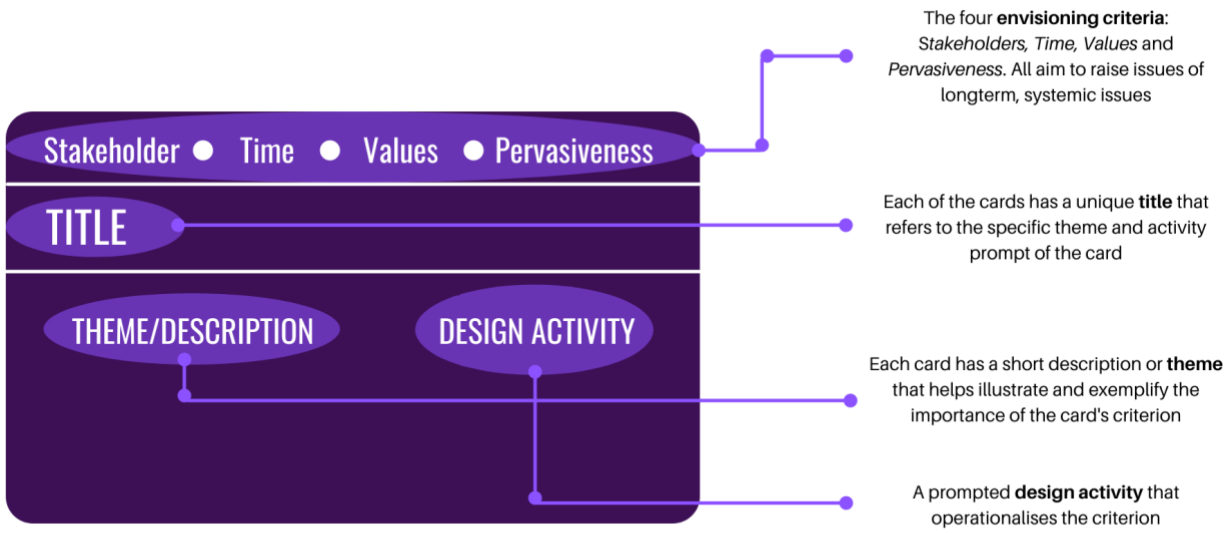

Figure 7. Envisioning Card (Source: Umbrello, n.d.)

The VSD approach to technology design has not remained solely within the realm of applications for contemporary technologies. It has gone beyond amelioration of problematic existent technologies,

such as energy infrastructure (Mouter et al., 2018; Oosterlaken, 2015), to speculative future technologies. Many speculative future technologies, including nanotechnology (Timmermans et al., 2011; Umbrello, 2019) and artificial intelligence (Umbrello and De Bellis, 2018; Umbrello and van de Poel, 2021), have received an abundance of attention from technical specialists and philosophers. Attention has also fallen on predictions for the sociocultural and ethical impacts these technologies will have. Current, more restricted forms of these technologies are already creating ethical tensions. For this reason, the imperative to guide their development towards beneficial ends has also garnered substantial attention (e.g., King et al., 2019; Umbrello and Baum, 2018).

For example, Timmermans et al. (2011) take nanopharmacy as their object of study. Nanopharmacy is the use of nanotechnology and nanomaterials for pharmaceutical applications. Although some nanomaterials and technologies are already employed in medicine, the technologies addressed here are speculative: Lab-on-a-Chip or Doctor-in-a-Cell technologies. The former are miniature metric devices that can be implanted in a patient to give accurate, real-time, tailored measurements of variables in relation to blood cells and genomes. Similarly, Doctor-in-a-Cell technologies are miniature interventions acting as "a molecular medical team that can be injected into a patient, coursing through his bloodstream [to] treat [medical problems]" (Casci, 2004; Timmermans et al., 2011). Like most nano, bio, information, and cognitive sciences (NBIC) technologies, they exist across multiple dimensions and blur the line of discrete practices. Medicine becomes intermingled even more closely with information and communications technology, which means values such as safety and privacy become just as blurred. Timmermans et al. propose a VSD approach to designing these technologies that envisions the various dimensions of values across disciplines, arguing that a VSD approach is necessary for the attainment of robust and salient design.

In a paper published in the *International Journal of Technoethics*, I likewise argue that VSD methodology is particularly potent for designing safe 'atomically precise' manufacturing (also known as molecular manufacturing or molecular fabricators) (Umbrello, 2019). Given the potential harms that advanced nanotechnology can have, which I describe elsewhere (Umbrello and Baum, 2018), I argue that the VSD approach can be adopted in current nanotechnology design practices. Adoption occurs by designing for four specific principles: proportionality, security, safety, and accessibility. While 'atomically precise' manufacturing is highly speculative, precursor technologies that provide the scaffolding for such technological advances can take these values into account. Decisions about physical and digital archetecture for current iterations of scaffolding technologies create elements of support and constraint for subsequent technological developments. To this end, VSD should not be

understood as discretely applicable towards only future 'atomically precise' manufacturing technologies. The VSD approach also applies to technologies that form its baser elements.

The same goes for artificial intelligence technolgies, which bear closer consideration here given the subject of this dissertation. In the edited collection *Artificial Intelligence Safety and Security* (2018), both Angelo de Bellis and myself published an article proposing the VSD approach as a particularly apt methodology in intelligent agent design (Umbrello and De Bellis, 2018). Whereas Aimee van Wynsberghe applied a modified version of VSD to care robots in her doctoral dissertation (Van Wynsberghe, 2013), we took her example in a more general application to increase the adoption of the VSD approach given the urgency of *in progressus* AI technologies. Arguing that current practices in AI development tend to be isolated to the decisions of designers and associated organisations, closed off from stakeholder enrollment, we call for a more situated and contextually based design approach to AI systems. Such systems have a distributed impact beyond direct stakeholders (designers and corporations). Valuation thus shifts away from a solely economic determinant towards a more harmonised amalgamation of human values such as privacy, safety, autonomy, and justice.

As mentioned, Aimee van Wynsberghe investigated the potential for integrating ethics into care robotics via implementation of a tailored form of VSD that she calls Care Centered Value Sensitive Design (CCVSD). The normative basis for her research shows how this tailored form of VSD is specific to the health and care sector by drawing upon traditional values that currently characterise healthcare (Van Wynsberghe, 2013). Her application illustrates how the VSD approach can be integrated into existing and developing frameworks. Successful integration means accounting for existing values, which she does. In doing so, she provides a means through which the methodology might be expanded outwards. She then accomplishes this by investigating the applicability of the CCVSD methodology to robots outside the healthcare domain, pushing the boundaries of the methodology to ascertain the limits of its application (Van Wynsberghe, 2016).

2.3.1 Commitments and Tensions

Despite the intentions and outcomes (discussed below) of the approach, VSD has not been without its critiques. Perhaps most notably, Davis and Nathan (2014) survey criticism from a variety of sources and divide them into four distinct categories that critique the VSD assumption of universal values, its ethical commitments, its stakeholder participation, and the voices of researchers and participants.

The traditional formulation of a VSD approach takes the universality of human values as a fundamental premise. Although it was never explicitly formulated in such a way, the function of VSD allows for reduction in the instantiation of values (or reduces stakeholder conceptions into concrete 'Western' valuations) (Borning and Muller, 2012). Of course, this is a contentious position in both philosophical and anthropological traditions (among others). It remains a contentious position within the VSD and design communities as a whole. I myself provided a substantial critique of the universalist position and its potential misgivings, and offered more practical heuristic tools to help designers avoid cognitive biases that are often exacerbated by transformative technologies (Umbrello, 2018). More recently, however, I provided a philosophical critique of the VSD assumption of universal values by arguing against this tradition and in favor of moral imagination theory (Johnson, 1993; Lakoff and Johnson, 2003). Moral imagination theory is a more appropriate moral landscape on which to practice VSD, enabling the cultivation of culturally sensitive design requirements (my paper is reprinted in full in Part II).

Because VSD does not prescribe any particular ethical theory while affirming universal values, it is open to particularly egregious moral theories such as Nazism (Albrechtslund, 2007). The founders argue that the instantiation of various values may be supported by some ethical theories (utilitarianism or deontological ethics), but not others (virtue ethical theories) (Friedman and Kahn Jr., 2003). This remains a point of contention and debate within VSD discourse. As I discuss in the second part of this dissertation, the debate results from VSD reliance on overly reductionist moral theories such as deontology, virtue ethics, or utilitarianism.

Scholars have also criticised stakeholder elicitation and participation, along with sufficient representation of stakeholder and designer voice throughout the design process, as a fundamental element of the approach. The *a priori* determination of relevant stakeholders, whether direct and indirect, has been subject to critique (Davis and Nathan, 2014; Manders-Huits, 2011). So, too, has the emergence of values from those elicitations (Borning and Muller, 2012; Le Dantec et al., 2009). Determining the identity of stakeholders for technologies that are often nebulous and cross domain is difficult. Likewise, it is challenging to identify technical means that elicit populations adequately enough to ensure an overall design that takes all affected stakeholders into sufficient consideration. To be fair, heuristic tools have been proposed and implemented with the aim of widening the scope of stakeholder analysis and range of enrollment; tools such as Envisioning Cards™ are intended to de-bias those elicitations (Friedman et al., 2017b; Friedman and Hendry, 2012). Other heuristic tools have been developed to de-bias the cognition of both designers and stakeholders during empirical elicitations

(Umbrello, 2018). Yoo (2017) develops stakeholder tokens, for instance, to identify impacted groups and relationships between stakeholders that may have been overlooked. This is a promising avenue for envisioning marginalised or otherwise overlooked stakeholders both early on and throughout design programs.

The same might be said for the voices of researchers and participants, as well as how a design program considers their expression. Borning and Muller (2012) hold a view similar to that of the anthropologist Donna Haraway, which argues that the position of the researcher as a disembodied unit (i.e., the gods-eye perspective) must be criticised for its more hegemonic tendencies in making design decisions. Instead, the designers themselves should be seen as a dynamic entity as much as any other stakeholder in the design program. My critique of the moral commitments of VSD in *Imaginative Value Sensitive Design: Using Moral Imagination Theory to Inform Responsible Technology Design* (Umbrello, 2020b) suggests a situated approach to determining how the designer implicates themselves as a stakeholder and influences design decisions (beyond simply imputing requirements elicited from stakeholder populations).

Finally, in their most recent work on VSD, Friedman and Hendry acknowledge criticism from the last two decades as well as developments towards a more holistic account of VSD and its methods. Ultimately, they leave most of the work on moral theorising and ethical commitments to future research avenues for scholars to tangle out. To some extent, my own work has aimed to do that (and I hope this dissertation further contributes to that objective). What moral theory, if any, should provide the foundation for the VSD approach? How do we sufficiently incorporate all stakeholders who are impacted by the technology, and how do we form complete population sets of those stakeholders? How does VSD impact (or how is it impacted by) policy and policy innovation? Given the ecological crises we are facing, which I discussed in greater depth in Umbrello (2018b), how do we account for the values of nonhuman (i.e., animal) stakeholders and the environment? All of these questions remain unanswered but deserve attention.

In sum, design teams have successfully adopted the VSD approach for existing technologies as well as the development and deployment of new iterations, such as web browsers and energy technologies. It has also been proposed as the means by which more speculative technology innovations, such as advanced nanotechnology and artificial intelligence systems, can be designed so as to map onto human values (rather than designing human values *ex post facto*). This does not mean that VSD is without misgivings or critiques. Whether in terms of its commitment to universal values, its privation of commitment to a single or comprehensive moral theory, its lack of guidance in determining

the most comprehensive set of relevant stakeholders, or its inability to ensure that their voices are adequately considered throughout the design phases remain marked points of research interest. Each of these points is of great importance to the efficacy and adoptability of VSD.

2.4 Research Outcomes

As discussed above, the VSD approach has been developed theoretically from philosophical and methodological perspectives. This has unfolded both conceptually and through practice over the last two decades since its conception. Scholarly literature has explored VSD applications ranging from real-world contemporary technologies such as web browser cookies (Millett et al., 2001) and IT systems for customs agencies (Vermaas et al., 2010) to more speculative technologies such as autonomous agents (Umbrello and De Bellis, 2018; Umbrello and van de Poel, 2021) and nanotechnology (Umbrello, 2019). However, the practice of VSD was never intended to only apply to early stages and during the design program. To ensure compliance and the ability of the artifact to respond to emerging values, it must also be applied post-deployment. To this end, I conclude the literature review in this subsection with illustrations of sustainable results from VSD programs. The remaining paragraphs describe an example of an artifact that was developed using the VSD approach, detailing its successes and challenges post-deployment.

Perhaps the longest lasting product developed from a VSD approach is UrbanSim, a "simulation system that models the development of urban areas over periods of twenty or more years" (Borning et al., 2004). The system is described on its website as software that

> leverages state-of-the-art urban simulation, 3D visualisation, and shared open data to empower users to explore, gain insights into, and develop and evaluate alternative plans to improve their communities. UrbanSim is a simulation platform for supporting planning and analysis of urban development, incorporating the interactions between land use, transportation, the economy, and the environment. ("UrbanSim," n.d.).

Created by Paul Waddell at the University of Washington, Borning et al. (2004) aided in the development of the technology by adding a participatory tool between designers and stakeholders. This tool determines the long-term impacts and alternatives in urban design projects. Using the VSD approach, Davis et al. (2006) formulated a list of goals to direct the design of UrbanSim:

- Improve system functionality by developing new tools for stakeholders to learn about, select, and visualise indicators to use in decision making.
- Support citizens and other stakeholders in evaluating alternatives with respect to their own values.

- Enhance system transparency with respect to its design, assumptions, and limitations – so it is not a black box.
- Contribute to system legitimacy by providing information that is credible and appropriate to the context for use.
- Foster citizen engagement in the decision-making process by providing tailored information and opportunities for involvement.

At the time of research and during early deployment of the software, its success spread out across five regions in the United States: Eugene/Springfield, Oregon; Honolulu, Hawaii; Salt Lake City, Utah; Houston, Texas; and the Puget Sound region, Washington (Borning et al., 2004). To date, adoption has spread across three continents and four additional countries encompassing Vancouver, Canada; Paris, France; and Johannesburg, South Africa. As a consequence, "over 51.7 million people live in areas covered by regional plans informed by UrbanCanvas Modeler and over 81.8 million people live in areas covered by regional plans informed by UrbanSim" ("UrbanSim," n.d.).

Although only a single illustration as to VSD efficacy and distribution has been provided here, it captures the consequences of VSD use: from its initial spatio-temporal location, the technology has since been distributed across sociocultural boundaries. For this reason, technological design must be fundamentally predicated on stakeholder voice and participation. Designing artifacts with those elicited values is critical. UrbanSim is a prime example of how such technologies construct lived environments. As such, they influence what comes after, how people interact, and how those environments will support or constrain the values of those who are bound to them. The continued adoption and distribution of UrbanSim is a testament to its adoptability and salience among stakeholders who participated in designing their urban environments – and, as a consequence, the approach used to design such a system.

3 Conclusion

More traditional design programs seek to create technologies by positioning economic values, such as efficiency and productivity, as the primary vectors. Value sensitive design moves away from this traditional conception to re-center the human as the vector from which design should emerge. This puts ecological and human values (even those of future generations) at the center of design programs rather than circumscribing them to *ad hoc* and *post hoc* additions. This literature review aims to distill foundational works in VSD to provide the reader with a thorough (albeit non-exhaustive) guide to how the VSD approach developed, its theoretical and methodological underpinnings, the projects that have adopted the approach, and how successful they turned out to be.

Appendix I

1. Betz, S., & Fritsch, A. (2016). A Comparison of Value Sensitive Design and Sustainability Design. In *Computer Science Spectrum* (pp. 267–274). Bonn: Society for Computer Science eV. https://subs.emis.de/LNI/Proceedings/Proceedings259/267.pdf%0Ahttp://cs.emis.de/LNI/Proceedings/Proceedings259/267.pdf
2. Briggs, P., & Thomas, L. (2015). An Inclusive, Value Sensitive Design Perspective on Future Identity Technologies. *ACM Transactions on Computer-Human Interaction*, *22*(5), 1–28. https://doi.org/10.1145/2778972
3. Cummings, M. L. (2006). Integrating ethics in design through the value-sensitive design approach. *Science and Engineering Ethics*, *12*(4), 701–715. https://doi.org/10.1007/s11948-006-0065-0
4. Davis, J., & Nathan, L. P. (2014). Value Sensitive Design: Applications, Adaptations, and Critiques. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* (pp. 1–26). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6994-6_3-1
5. Friedman, B. (1996). Value-sensitive design. *Interactions*, *3*(6), 16–23. https://doi.org/10.1145/242485.242493
6. Friedman, B. (1999). Value-sensitive design: A research agenda for information technology. *Contract No: SBR-9729633). National Science Foundation, Arlington, VA*.
7. Friedman, B., Howe, D. C., & Felten, E. (2002). Informed consent in the Mozilla browser: Implementing value-sensitive design. In *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on* (pp. 10-pp). IEEE.
8. Friedman, B., & Kahn Jr., P. H. (2002). Value sensitive design: Theory and methods. *University of Washington Technical*, (December), 1–8. https://doi.org/10.1016/j.neuropharm.2007.08.009
9. Friedman, B., Kahn Jr., P. H., Borning, A., & Huldtgren, A. (2013). Value Sensitive Design and Information Systems. In N. Doorn, D. Schuurbiers, I. van de Poel, & M. E. Gorman (Eds.), *Early engagement and new technologies: Opening up the laboratory* (pp. 55–95). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-7844-3_4
10. 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
11. Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. Cambridge, MA: Mit Press.
12. Manders-Huits, N. (2011). What Values in Design? The Challenge of Incorporating Moral Values into Design. *Science and Engineering Ethics*, *17*(2), 271–287. https://doi.org/10.1007/s11948-010-9198-2
13. Vermaas, P. E., Tan, Y.-H., van den Hoven, J., Burgemeestre, B., & Hulstijn, J. (2010). Designing for trust: A case of value-sensitive design. *Knowledge, Technology & Policy*, *23*(3–4), 491–505.
14. Winkler, T., & Spiekermann, S. (2018). Twenty years of value sensitive design: a review of methodological practices in VSD projects. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-018-9476-2
15. Yoo, D., Huldtgren, A., Woelfer, J.P., Hendry, D.G., Friedman, B., 2013. A value sensitive action-reflection model: evolving a co-design space with stakeholder and designer prompts, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 419–428.

Appendix II

"query": { value sensitive design }
"filter": { ACM Pub type: Journals, Publication Date: (01/01/1996 TO 12/31/2020), ACM Content: DL }

"query": { value-sensitive design }
"filter": { ACM Pub type: Journals, Publication Date: (01/01/1996 TO 12/31/2020), ACM Content: DL }

"query": { vsd }

"filter": { ACM Pub type: Journals, Publication Date: (01/01/1996 TO 12/31/2020), ACM Content: DL }

References

Ackerman, M.S., Cranor, L., 1999. Privacy Critics: UI Components to Safeguard Users' Privacy, in: CHI '99 Extended Abstracts on Human Factors in Computing Systems, CHI EA '99. ACM, New York, NY, USA, pp. 258–259. https://doi.org/10.1145/632716.632875
Agre, P.E., 1997. Beyond the Mirror World: Privacy and the Representational Practices of Computing. Technology and Privacy: The New Landscape. PE Agre and M. Rotenberg.
Albrechtslund, A., 2007. Ethics and technology design. Ethics Inf. Technol. 9, 63–72.
Borning, A., Friedman, B., Kahn, P., 2004. Designing for human values in an urban simulation system: Value sensitive design and participatory design, in: Proceedings From the Eighth Biennial Participatory Design Conference.
Borning, A., Muller, M., 2012. Next steps for value sensitive design. Proc. 2012 ACM Annu. Conf. Hum. Factors Comput. Syst. - CHI '12 1125. https://doi.org/10.1145/2207676.2208560
Casci, T., 2004. Doctor in a cell. Nat. Rev. Genet. 5, 406.
Cooper, H.M., 1988. Organizing knowledge syntheses: A taxonomy of literature reviews. Knowl. Soc. 1, 104.
Cummings, M.L., 2006. Integrating ethics in design through the value-sensitive design approach. Sci. Eng. Ethics 12, 701–715. https://doi.org/10.1007/s11948-006-0065-0
Davis, J., Lin, P., Borning, A., Friedman, B., Kahn, P.H., Waddell, P.A., 2006. Simulations for urban planning: Designing for human values. Computer (Long. Beach. Calif). 39, 66–72.
Davis, J., Nathan, L.P., 2015. Handbook of ethics, values, and technological design: Sources, theory, values and application domains, in: van den Hoven, J., Vermaas, P.E., van de Poel, I. (Eds.), Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains. pp. 12–40. https://doi.org/10.1007/978-94-007-6970-0
Davis, J., Nathan, L.P., 2014. Value Sensitive Design: Applications, Adaptations, and Critiques, in: van den Hoven, J., Vermaas, P.E., van de Poel, I. (Eds.), Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains. Springer Netherlands, Dordrecht, pp. 1–26. https://doi.org/10.1007/978-94-007-6994-6_3-1
Fogg, B.J., Tseng, H., 1999. The elements of computer credibility, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 80–87.
Friedman, B., 1999. Value-sensitive design: A research agenda for information technology. Contract No SBR-9729633). Natl. Sci. Found. Arlington, VA.
Friedman, B., 1996. Value-sensitive design. Interactions 3, 16–23. https://doi.org/10.1145/242485.242493
Friedman, B., Hendry, D.G., 2019. Value Sensitive Design: Shaping Technology with Moral Imagination. Mit Press, Cambridge, MA.
Friedman, B., Hendry, D.G., 2012. The envisioning cards: A toolkit for catalyzing humanistic and technical imaginations, in: Proceedings of the 30th International Conference on Human Factors in Computing Systems - CHI '12. pp. 1145–1148. https://doi.org/10.1145/2207676.2208562
Friedman, B., Hendry, D.G., Borning, A., 2017a. A Survey of Value Sensitive Design Methods. Found. Trends® Human–Computer Interact. 11, 63–125. https://doi.org/10.1561/1100000015
Friedman, B., Hendry, D.G., Huldtgren, A., Jonker, C., Van den Hoven, J., Van Wynsberghe, A., 2015. Charting the Next Decade for Value Sensitive Design. Aarhus Ser. Hum. Centered Comput. 1, 4. https://doi.org/10.7146/aahcc.v1i1.21619
Friedman, B., Howe, D.C., Felten, E., 2002. Informed consent in the Mozilla browser: Implementing value-sensitive design, in: System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference On. IEEE, pp. 10-pp.
Friedman, B., Kahn Jr., P.H., 2003. Human values, ethics, and design, in: Jacko, J.A., Sears, A. (Eds.), The Human-Computer Interaction Handbook. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, pp. 1177–1201.

Friedman, B., Kahn Jr., P.H., 2002. Value sensitive design: Theory and methods. Univ. Washingt. Tech. 1–8. https://doi.org/10.1016/j.neuropharm.2007.08.009
Friedman, B., Kahn Jr., P.H., Borning, A., 2008. Value Sensitive Design and Information Systems. Human-Computer Interact. Manag. Inf. Syst. Found. 69–101. https://doi.org/10.1145/242485.242493
Friedman, B., Kahn Jr., P.H., Borning, A., Huldtgren, A., 2013a. Value Sensitive Design and Information Systems, in: Doorn, N., Schuurbiers, D., van de Poel, I., Gorman, M.E. (Eds.), Early Engagement and New Technologies: Opening up the Laboratory. Springer Netherlands, Dordrecht, pp. 55–95. https://doi.org/10.1007/978-94-007-7844-3_4
Friedman, B., Kahn Jr, P.H., 2007. Human values, ethics, and design, in: The Human-Computer Interaction Handbook. CRC Press, pp. 1223–1248.
Friedman, B., Kahn, P.H., Borning, A., Huldtgren, A., 2013b. Value Sensitive Design and Information Systems, in: Doorn, N., Schuurbiers, D., van de Poel, I., Gorman, M.E. (Eds.), Early Engagement and New Technologies: Opening up the Laboratory. Springer Netherlands, Dordrecht, pp. 55–95. https://doi.org/10.1007/978-94-007-7844-3_4
Friedman, B., Nathan, L.P., Kane, S.K., Lin, J., 2017b. Envisioning Cards.
Friedman, B., Nissenbaum, H., 1996. Bias in computer systems. ACM Trans. Inf. Syst. 14, 330–347.
Fuchs, L., 1999. AREA: a cross-application notification service for groupware, in: ECSCW'99. Springer, pp. 61–80.
Gazzaneo, L., Padovano, A., Umbrello, S., 2020. Designing Smart Operator 4.0 for Human Values: A Value Sensitive Design Approach. Procedia Manuf. 42, 219–226. https://doi.org/10.1016/j.promfg.2020.02.073
IEEE, 2019. Ethically Aligned Design, IEEE Standards v2.
Jancke, G., Venolia, G.D., Grudin, J., Cadiz, J.J., Gupta, A., 2001. Linking public spaces: technical and social issues, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 530–537.
Johnson, M., 1993. Moral Imagination: Implications of Cognitive Science for Ethics. University of Chicago Press, Chicago, IL.
King, T.C., Aggarwal, N., Taddeo, M., Floridi, L., 2019. Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. Sci. Eng. Ethics. https://doi.org/10.1007/s11948-018-00081-0
Lakoff, G., Johnson, M., 2003. Metaphors We Live By. University of Chicago Press, Chicago, IL.
Le Dantec, C.A., Poole, E.S., Wyche, S.P., 2009. Values As Lived Experience: Evolving Value Sensitive Design in Support of Value Discovery, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09. ACM, New York, NY, USA, pp. 1141–1150. https://doi.org/10.1145/1518701.1518875
Leveson, N.G., 1991. Software safety in embedded computer systems. Commun. ACM 34, 34–46.
Lipinski, T.A., Britz, J., 2000. Rethinking the ownership of information in the 21st century: Ethical implications. Ethics Inf. Technol. 2, 49–71.
Manders-Huits, N., 2011. What Values in Design? The Challenge of Incorporating Moral Values into Design. Sci. Eng. Ethics 17, 271–287. https://doi.org/10.1007/s11948-010-9198-2
Millett, L.I., Friedman, B., Felten, E., 2001. Cookies and web browser design: toward realizing informed consent online, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 46–52.
Mouter, N., de Geest, A., Doorn, N., 2018. A values-based approach to energy controversies: Value-sensitive design applied to the Groningen gas controversy in the Netherlands. Energy Policy 122, 639–648.
Oosterlaken, I., 2015. Applying Value Sensitive Design (VSD) to Wind Turbines and Wind Parks: An

Exploration. Sci. Eng. Ethics 21, 359–379. https://doi.org/10.1007/s11948-014-9536-x
Palen, L., Grudin, J., 2003. Discretionary adoption of group support software: Lessons from calendar applications, in: Implementing Collaboration Technologies in Industry. Springer, pp. 159–180.
Randolph, J., 2009. A guide to writing the dissertation literature review. Pract. Assessment, Res. Eval. 14, 13.
Rocco, E., 1998. Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 496–502.
Shneiderman, B., 1999. Universal usability: pushing human–computer interaction research to empower every citizen. ISRTechnical Report, in: University of Maryland, Institute for Systems Research, College Park. Citeseer.
Suchman, L., 1993. Do categories have politics? The language/action perspective reconsidered, in: Proceedings of the Third European Conference on Computer-Supported Cooperative Work 13–17 September 1993, Milan, Italy ECSCW'93. Springer, pp. 1–14.
Tang, J.C., 1997. Eliminating a hardware switch: weighing economics and values in a design decision, in: Human Values and the Design of Computer Technology. Center for the Study of Language and Information, pp. 259–269.
Thomas, J.C., 1997. Steps toward universal access within a communications company. Hum. Values Des. Comput. Technol. Cambridge Univ. Press. New York, NY 271–287.
Timmermans, J., Zhao, Y., van den Hoven, J., 2011. Ethics and Nanopharmacy: Value Sensitive Design of New Drugs. Nanoethics 5, 269–283. https://doi.org/10.1007/s11569-011-0135-x
Umbrello, S., 2020a. Meaningful Human Control over Smart Home Systems: A Value Sensitive Design Approach. Humana.Mente J. Philos. Stud. 13, 40–65.
Umbrello, S., 2020b. Imaginative Value Sensitive Design: Using Moral Imagination Theory to Inform Responsible Technology Design. Sci. Eng. Ethics 26, 575–595. https://doi.org/10.1007/s11948-019-00104-4
Umbrello, S. (2021). Conceptualizing Policy in Value Sensitive Design: A Machine Ethics Approach. In S. J. Thompson (Ed.), *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence* (pp. 108–125). IGI Global. https://doi.org/10.4018/978-1-7998-4894-3.ch007
Umbrello, S., 2019. Atomically Precise Manufacturing and Responsible Innovation: A Value Sensitive Design Approach to Explorative Nanophilosophy. Int. J. Technoethics 10, 1–21. https://doi.org/10.4018/IJT.2019070101
Umbrello, S., 2018. The moral psychology of value sensitive design: the methodological issues of moral intuitions for responsible innovation. J. Responsible Innov. 5, 186–200. https://doi.org/10.1080/23299460.2018.1457401
Umbrello, S., n.d. The Role of Engineers in Harmonizing Human Values for AI Systems Design. *Working paper.*
Umbrello, S., Baum, S.D., 2018. Evaluating future nanotechnology: The net societal impacts of atomically precise manufacturing. Futures 100, 63–73. https://doi.org/10.1016/j.futures.2018.04.007
Umbrello, S., De Bellis, A.F., 2018. A Value-Sensitive Design Approach to Intelligent Agents, in: Yampolskiy, R. V. (Ed.), Artificial Intelligence Safety and Security. CRC Press, pp. 395–410. https://doi.org/10.13140/RG.2.2.17162.77762
Umbrello, S., van de Poel, I., 2021. Mapping Value Sensitive Design onto AI for Social Good Principles. AI and Ethics. https://doi.org/10.1007/s43681-021-00038-3
UrbanSim [WWW Document], n.d. URL https://urbansim.com/home (accessed 3.16.20).
van de Poel, I., 2014a. Design for Values in Engineering, in: van den Hoven, J., Vermaas, P.E., van de Poel, I. (Eds.), Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values

and Application Domains. Springer Netherlands, Dordrecht, pp. 1–20. https://doi.org/10.1007/978-94-007-6994-6_25-1
van de Poel, I., 2014b. Conflicting Values in Design, in: van den Hoven, J., Vermaas, P.E., van de Poel, I. (Eds.), Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains. Springer Netherlands, Dordrecht, pp. 1–23. https://doi.org/10.1007/978-94-007-6994-6_5-1
van den Hoven, J., Lokhorst, G.J., van de Poel, I., 2012. Engineering and the Problem of Moral Overload. Sci. Eng. Ethics 18, 143–155. https://doi.org/10.1007/s11948-011-9277-z
van den Hoven, J., Manders-Huits, N., 2009. Value-Sensitive Design, in: A Companion to the Philosophy of Technology. Wiley-Blackwell, pp. 477–480. https://doi.org/10.1002/9781444310795.ch86
van den Hoven, J., Vermaas, P.E., van de Poel, I., 2015. Handbook of ethics, values, and technological design: Sources, theory, values and application domains, Springer Reference. Springer Netherlands. https://doi.org/10.1007/978-94-007-6970-0
van den Hoven, J., Weckert, J., 2008. Information Technology and Moral Philosophy. Cambridge University Press.
van Wynsberghe, A., 2016. Service robots, care ethics, and design. Ethics Inf. Technol. 18, 311–321. https://doi.org/10.1007/s10676-016-9409-x
van Wynsberghe, A., 2013. Designing Robots for Care: Care Centered Value-Sensitive Design. Sci. Eng. Ethics 19, 407–433. https://doi.org/10.1007/s11948-011-9343-6
Vermaas, P.E., Tan, Y.-H., van den Hoven, J., Burgemeestre, B., Hulstijn, J., 2010. Designing for Trust: A Case of Value-Sensitive Design. Knowledge, Technol. Policy 23, 491–505. https://doi.org/10.1007/s12130-010-9130-8
Winkler, T., Spiekermann, S., 2018. Twenty years of value sensitive design: a review of methodological practices in VSD projects. Ethics Inf. Technol. https://doi.org/10.1007/s10676-018-9476-2
Winograd, T., 1993. Categories, disciplines, and social coordination. Comput. Support. Coop. Work 2, 191–197.
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. ACM, pp. 280–284.
Yoo, D., Huldtgren, A., Woelfer, J.P., Hendry, D.G., Friedman, B., 2013. A value sensitive action-reflection model: evolving a co-design space with stakeholder and designer prompts, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 419–428.
Zheng, J., Bos, N., Olson, J.S., Olson, G.M., 2001. Trust with out touch: jump-start trust with social chat, in: CHI'01 Extended Abstracts on Human Factors in Computing Systems. pp. 293–294.

PART I

A PHILOSOPHY OF SYSTEMS THINKING AND MEANINGFUL HUMAN CONTROL

Ethics and Information Technology
https://doi.org/10.1007/s10676-021-09588-w

ORIGINAL PAPER

Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: a two-tiered approach

Steven Umbrello[1]

Accepted: 29 March 2021

Abstract
The international debate on the ethics and legality of autonomous weapon systems (AWS), along with the call for a ban, primarily focus on the nebulous concept of fully autonomous AWS. These are AWS capable of target selection and engagement absent human supervision or control. This paper argues that such a conception of autonomy is divorced from both military planning and decision-making operations; it also ignores the design requirements that govern AWS engineering and the subsequent tracking and tracing of moral responsibility. To show how military operations can be coupled with design ethics, this paper marries two different kinds of meaningful human control (MHC) termed levels of abstraction. Under this two-tiered understanding of MHC, the contentious notion of 'full' autonomy becomes unproblematic.

Keywords Meaningful human control · Autonomous weapons · Systems theory · Design for values · Applied ethics

Introduction

Although technological innovations have always played a key role in military operations, autonomous weapons systems (AWS) have received asymmetric attention in public debate as well as academic discussions—and for good reason (Kania, 2017). As these systems are designed to carry out more and more tasks once in the domain of human operators, questions regarding their autonomy and potential recalcitrance have sparked discussion. Debate highlights a potential *accountability gap* between their use and who, if anyone, can be held accountable. At the international level, discussions about how to exercise control over the development and deployment of these autonomous military systems have been underway for over a decade. Still, there remains very little consensus as to what constitutes a sufficient level of control.

The concept of *meaningful human control* (MHC) has emerged in discourse to encompass this ideal of human control over autonomous systems. Various approaches have been taken to define a sufficiently robust notion of MHC that addresses technical requirements (Arkin, 2008), proper training for use (Article36, 2015; Asaro, 2009), designer-user engagement (Leveringhaus, 2016), operations planning (Ekelhof, 2019), design requirements, and the responsibility of designers (Elands et al., 2019; Mecacci and Santoni de Sio 2019; Santoni de Sio and Van den Hoven 2018). Each of these approaches provide insight into how MHC over these types of systems can be understood and attained. Although they are generally proposed as isolated frameworks for attaining MHC, they share some underlying precepts. Approaches that emphasize the operational planning and military context of use, such as that of Ekelhof (2019), provide a strong contextual landscape for understanding MHC. Other approaches, such as that of Santoni de Sio et al. (2018, 2019), focus on design histories, designer intentions and plans, or the responsibilities of designers and suprain-dividual agents. They provide cogent arguments for designing these systems with both backward- and forward-looking responsibility. Still, they largely focus on a single level of abstraction at the opportunity cost of the other.

This paper aims to employ the concepts of systems theory (theoretical lens) and systems engineering (applied lens) to understand MHC across these levels of abstraction (LoA). As a way of marrying these often-isolated projects of defining MHC, I propose a two-tiered approach to understanding MHC. First, it does not make sense to divorce discussions of AWS from actual and often trivial military operations; AWS exist within this landscape, not outside it. One must

✉ Steven Umbrello
steven.umbrello@unito.it

[1] Institute for Ethics and Emerging Technologies, University of Turin, Via Sant'Ottavio, 20, 10124 Torino, TO, Italy

Published online: 05 April 2021

2 Systems Theory: An Ontology *for* Engineering

Systems Thinking is a mixed bag of holistic, balanced and often abstract thinking to understand things profoundly and solve problems systematically — Pearl Zhu

2.1 Introduction

Although technological innovations have always played a key role in military operations, autonomous weapons systems (AWS) are receiving asymmetric attention both in public debate as well as academic discussions – and for good reason (Kania, 2017). These systems are designed to carry out tasks that were once exclusive to the domain of human operators. Questions regarding their autonomy and potential recalcitrance have sparked discussions that highlight a potential *accountability gap* between their use and who, if anyone, should be held accountable. At an international level, discussions regarding how to exercise control over the development and deployment of these autonomous military systems have been ongoing for over a decade. There remains very little consensus as to what constitutes a sufficient level of control.

In this debate, the concept of *meaningful human control* (MHC) emerged to encompass an ideal of human control over autonomous systems. Various approaches have been taken to define a sufficiently complete notion of MHC that ranges from technical requirements (Arkin, 2008), proper training for use (Article 36, 2015; Asaro, 2009), designer-user engagement (Leveringhaus, 2016), and operations planning (Ekelhof, 2019) to design requirements and the responsibility of designers (Mecacci & de Sio, 2019; Santoni de Sio & van den Hoven, 2018). Each of these approaches provides insight into how to attain or understand MHC over these types of systems. Although they are generally seen as isolated frameworks for attaining MHC, they all share some underlying precepts. Approaches that emphasise operational planning and the military context for use, as applied by Ekelhof (2019), provide a strong contextual landscape for understanding MHC. Other approaches that focus on design histories, the intentions and plans of designers, or the responsibility of designers and supraindividual agents, as described in Santoni de Sio et al. (2018;2019), provide cogent arguments for designing these systems with backward- and forward-looking responsibility. Still, they largely focus on a single level of abstraction at the opportunity cost of other levels.

This chapter employs the concepts of systems theory as a theoretical lens, and systems engineering as an applied lens. Together, the two lenses provide an ontology for understanding MHC across these levels of abstraction. But the chapter also understands these concepts as a motivating

factor for adoption of the VSD approach to design methodology for MHC. In doing so, it provides a more coherent and explicit ontology of engineering. The following chapters will then use this ontology to construct a two-tiered approach to understanding MHC – one that marries these often-isolated projects to arrive at a definition. It does not make sense to divorce discussions of AWS from actual and often trivial military operations; AWS exist within this landscape, not outside of it. One must therefore situate these systems within their operational context (Operational Level) to understand AWS. This does not mean there are no accountability gaps with fully AWS. But when it comes to determining the responsiveness of a system to the relevant moral reasons of relevant agents, the design question is still important (Design Level). Part I outlines how the coupling of these levels of abstraction can account for technical full autonomy of certain types of AWS. This account not only resolves many of the issues regarding (fully) AWS, but actually provides the key to achieving MHC.[13]

2.2 Systems Thinking

2.2.1 Why an Ontology of Systems?

The term systems theory is *prima facie* self-explanatory. But a definition of its meaning merits mentioning *why* a dissertation conceptualising a theory of MHC and its application (i.e., design) warrants any discussion of more abstract ontology. There are multiple reasons for drawing an ontology. To begin, the primary reason for adopting systems theory as the ontological framework for this investigation is that it (implicitly) characterises the two levels of abstraction for understanding MHC discussed in the chapters that follow along with Annex I. The operational level of control is characterised by a plurality of actors and networks that complicates, yet also constitutes, how military operations are structured, planned, and carried out. Likewise, the design level of control is fundamentally built on the notion of tracking and tracing networks of systems and actors both in the use and in the design histories of those systems.

Secondly, systems theory is the theoretical framework from which systems engineering derives. As discussed in the next subsection, systems engineering developed in the domain of defence. It is essentially the practical and managerial implementation of a systems thinking[14] ontology. Aside from

[13] It should be noted that the argument forwarded by this chapter (and dissertation more broadly) does not advocate *for* the development of (fully) AWS. Rather, it focuses on the notion of control over certain types of AWS given how current military operations *actually* function as well as how design practices contribute to control. At the very least, this section aims to highlight a potential gap that theorists and policy-makers can address when formulating their own arguments on if/how AWS are ethically problematic and whether *certain* types of AWS should be prohibited.

[14] The term 'Systems Thinking' here is used in the verbial sense, that is, conceptualizing things in terms of systems, or, more poignantly, within the axioms of systems theory.

the obvious congruency between systems engineering and systems thinking within the military sphere, VSD exists as a sort of parallel approach to systems thinking design methodology. As discussed in more detail in Annex II, VSD is fundamentally predicated on a systems thinking approach to design. Affirming an *interactional* stance on technology, VSD acknowledges that technology and societal forces co-construct and co-vary with one another (Friedman et al., 2017). As a result, technology is neither purely deterministic nor instrumental –nor is society wholly constructivist. Rather, various actors, institutions, technologies, and their design histories form complex yet important networks of interaction. These relationships need to be brought to the fore for salient and responsible innovation to take place.

This is the substrata that underlies the coupling of two levels of abstraction for understanding MHC. Exploration of (fully) AWS and the use of VSD to design *for* MHC is necessary. Such exploration provides a landscape in which diverse moral universes from different societies and cultures (each with their own moral traditions and heritage) can come together in good faith for discourse on how to confront AWS. Here, systems thinking is the philosophical precept motivating most engineering programs across the globe along with their more specific military domains. It thus serves as the engineering Rosetta Stone for coupling the two levels of abstraction to understand MHC for fully AWS. Likewise, its substructure is the common thread unifying this conception of MHC and the VSD approach to designing *for* such control.

2.2.2 Organisation, Connection, and Complexity

Systems theory is broadly understood as an interdisciplinary study of organised and complex systems (Whitchurch and Constantine 2009). A system can be understood as a connected cluster of both co-constitutive and co-varying parts that may be synthetic and/or biological. Systems are understood as fundamentally constrained by spatiotemporal vectors, altered by their context or environment, and defined by their architecture and teleology (the latter of which is expressed through operation) (Adams et al. 2014). To this end, systems are often characterised as being *more* than the sum of their constituent parts if they express emergent behavior (Dudo et al. 2011; Wan 2011) or synergy (Haken 2013). Alteration at any given node(s) of a system can result in alteration at other node(s) as well as the resulting emergent behavior (if any). One of the aims of systems theory is to map out patterns of behavior for these complex systems to better predict future behavior based on environmental inputs.

This is particularly true for systems that adapt and learn (i.e., machine learning) from their environmental context (Ivanov 1993). Similarly, systems can both support and constrain other systems

to make them more or less robust. Systems theory generally seeks to understand the kinetics of systems, their pressures and conditions, and general methods and tools. These can be extrapolated to better understand other systems at all levels of recursion (Graham et al. 1994) across a variety of fields (i.e., biology, chemistry, ecology, engineering, and psychology) with the aim of optimising equifinality (Beven 2006).

General systems theory (GST) thus aims to develop tools and methods for a general understanding of complex systems rather than specific approaches to a single system or domain (Von Bertalanffy 1972). GST makes further distinctions between system types or, more specifically, between active systems and passive ones. Active systems are characterised by structures or components that engage in processes and exhibit active behaviour, while passive systems are those structures that are engaged or processed. An AWS is a passive system when it is powered down or lacks a power source; it is an active system when booted and deployed in the field. In other words, any given system can be both passive and active at any given spatiotemporal vector. Any given system can also be composed of both passive and active systems. This framing is particularly relevant to an ontological understanding of complex artificial intelligence (AI) systems, which employ what are often considered opaque algorithmic processes that result from hybrid machine learning and neural network systems like those being considered for use in AWS (Boscoe 2019; Turilli and Floridi 2009; Wachter et al. 2017). Given the complexity and need to direct optimal systems design, systems engineering becomes particularly relevant to the applied domains of this theory.

2.3 Systems Engineering

Systems engineering then takes the multidisciplinary approach to understanding systems and applies it to the understanding, design, management, and deployment of engineered systems to ensure optimised equifinality over their lifecycles (Adams et al. 2014; Thomé 1993). Engineered systems are designed in such a way as to ensure constituent parts work synergistically. When they do, emergent behaviours are beneficial. Additionally, systems engineering draws on many overlapping human-centric disciplines, such as risk analysis, organisational studies, and project management (i.e., paralleling the operations planning of Ekelhof's (2019) conception of MHC) as well as technical disciplines, such as requirements engineering, cybernetics, software and electrical engineering, and industrial engineering, among others. In doing so, it frames the engineering processes themselves holistically as part of the larger system that conditions the project being undertaken.

As mentioned above, this approach to conceptualising engineering practice originates from the defence industry. Since WWII, it has been in continuous (albeit continually morphing) use within the defence domain. This is mostly on account of the approach's performance history of mitigating reliability risks where proper systemic function is existential. A direct, proportional relationship between project performance and the application of systems engineering approaches was demonstrated in a collaborative study between Carnegie Mellon University, the Software Engineering Institute, the IEEE Aerospace and Electronic Systems Society, and the National Defense Industrial Association (Elm and Goldenson, 2012). Drawing from systems thinking, systems engineering aims to optimise equifinality by approaching the complexity of technologies as dynamic, continually changing systems that likewise require co-design and monitoring for their full life cycle (SyntheSys, 2020).

Full life cycle monitoring and meeting the needs of changing design requirements stems from the complexity of dynamic systems as a function of their emergent properties, and thus changing values. UK-based information systems engineering firm SyntheSys Technologies argues that approaching engineering this way "has produced a robust and scientific approach to requirements management and verification, a greater focus on the full life cycle of a product, and novel modelling techniques for complex emergent behaviour" (SyntheSys, 2020). At their core, systems engineering models are predicated on the precept that system performance is a consequence of system architecture. This means individual nodes, which constitute any given system's black box of 'system elements', form a system when assembled in an organised environment. These systems can be clustered into more complex networks wherein every given subsystem, which independently constitute a larger system of systems, nonetheless functions as a predicate of the emergent performance of the whole. This is illustrated in, for example, a military's information and communications technology network or its global logistics systems. Modelling systems this way allows designers to emphasise the nuanced interconnections that constitute a system along with the relationships and impacts of the system(s) in dynamic environments. It permits greater reflexivity to changing needs that result from emergent behaviours, thus helping to better predict or specify the complexity of causal relationships. As a consequence, it can address the unforeseen or even unforeseeable consequences of system performance *in situ*. This type of modelling, then, is particularly attractive for determining the potential opportunity costs of pursuing any given system architecture in any given environment. Once costs are evident, it allows for intervention to address any particular issue early on in the design process – thus avoiding unnecessary, wanton spending.

Likewise, coordination and management between modelling domains involved in systems engineering are itself part of the 'system in a system'. These layers are unified through uniform systems modelling languages such as SysMLTM, increasing the equifinality of the engineers within the system (OMG and SysMLTM 2017; SyntheSys, 2020). This permits more accurate integration, verification, and validation of system requirements across separate (albeit cooperating) engineering spaces and deployment environments.

2.4 Conclusions

On a pragmatic level, systems engineering involves anticipating client needs and specific design requirements early on in the development cycle. When this has been achieved, engineers can then move on to design synthesis and system validation while continually maintaining a holistic picture of the development life cycle of the system (i.e., systems *thinking*). In order to do this successfully, designers must consider all of the potentially implicated stakeholders and their values as pertains to the design project. This latter point on stakeholders is discussed in greater detail in Part II; it directly aligns with theories of responsible innovation and value sensitive design (VSD), in particular (Santoni di Sio et al. 2018 claim their conception of MHC arises from and aligns with VSD at a design level). By a similar token, systems thinking in general (i.e., systems theory + systems engineering) offers a reasonable tool for framing the common ground and need to combine the two levels of abstraction to formulate a similarly holistic understanding of MHC.

References

Adams, K. M., Hester, P. T., Bradley, J. M., Meyers, T. J., & Keating, C. B. (2014). Systems theory as the foundation for understanding systems. *Systems Engineering*, *17*(1), 112–123.

Arkin, R. C. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture part I: Motivation and philosophy. In *Proceedings of the 3rd international conference on Human robot interaction - HRI '08* (p. 121). New York, New York, USA: ACM Press. https://doi.org/10.1145/1349822.1349839

Article 36. (2015). Killing by machine: Key issues for understanding meaningful human control. Retrieved January 28, 2020, from http://www.article36.org/weapons/autonomous-weapons/killing-by-machine-key-issues-for-understanding-meaningful-human-control/

Asaro, P. (2009). Modeling the moral user. *IEEE Technology and Society Magazine*, *28*(1), 20–24. https://doi.org/10.1109/MTS.2009.931863

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of hydrology*, *320*(1–2), 18–36.

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

Dudo, A., Dunwoody, S., & Scheufele, D. A. (2011). The Emergence of Nano News: Tracking

Thematic Trends and Changes in U.S. Newspaper Coverage of Nanotechnology. *Journalism & Mass Communication Quarterly*, *88*(1), 55–75. https://doi.org/10.1177/107769901108800104
Elm, Joseph P., and Dennis R. Goldenson. 2012. "The Business Case for Systems Engineering Study: Results of the Systems Engineering Effectiveness Survey." https://www.ndia.org/-/media/sites/ndia/meetings-and-events/divisions/systems-engineering/studies-and-publications/business-case-for-se---results.ashx.
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
Friedman, Batya, David G. Hendry, and Alan Borning. 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.
Graham, R., Knuth, D., & Patashnik, O. (1994). 1. Recurrent Problems. In *Concrete Mathematics: A Foundation for Computer Science* (Second., p. 670). Reading, Massachusetts: Addison-Wesley Professional.
Haken, H. (2013). *Synergetics: Introduction and advanced topics*. Springer Science & Business Media.
Ivanov, K. (1993). Hypersystems: a base for specification of computer-supported self-learning social systems. In *Comprehensive systems design: A new educational technology* (pp. 381–407). Springer.
Kania, E. B. (2017). Battlefield Singularity. *Artificial Intelligence, Military Revolution, and China's Future Military Power, CNAS*.
Leveringhaus, A. (2016). Drones, automated targeting, and moral responsibility. In E. Di Nucci & F. Santoni de Sio (Eds.), *Drones and Responsibility: Legal, Philosophical, and Socio-Technical Perspectives on the Use of Remotely Controlled Weapons* (pp. 169–181). Routledge. https://doi.org/9781138390669
Mecacci, G., & de Sio, F. S. (2019). Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics and Information Technology*, 1–13.
OMG, and SysMLTM. 2017. "Systems Modeling Language." Version.
Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account . *Frontiers in Robotics and AI* . Retrieved from https://www.frontiersin.org/article/10.3389/frobt.2018.00015 SyntheSys. 2020. "Why Use
Systems Engineering?" *The IT Insider*, July 2020. https://theitinsider.co.uk/articles/2020/why-use-systems-engineering/.
Thomé, B. (1993). *Systems engineering: principles and practice of computer-based systems engineering*. John Wiley and Sons Ltd.
Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*(2), 105–112. https://doi.org/10.1007/s10676-009-9187-9
Von Bertalanffy, L. (1972). The history and status of general systems theory. *Academy of management journal*, *15*(4), 407–426.
Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics*, *2*(6), eaan6080.
Wan, P. Y. (2011). Emergence à la systems theory: epistemological Totalausschluss or ontological novelty? *Philosophy of the Social Sciences*, *41*(2), 178–210.
Whitchurch, G. G., & Constantine, L. L. (2009). Systems theory. In *Sourcebook of family theories and methods* (pp. 325–355). Springer.

Meaningful Human Control over Smart Home Systems: A Value Sensitive Design Approach

Steven Umbrello[†]
steven.umbrello@unito.it

ABSTRACT

The last decade has witnessed the mass distribution and adoption of smart home systems and devices powered by artificial intelligence systems ranging from household appliances like fridges and toasters to more background systems such as air and water quality controllers. The pervasiveness of these sociotechnical systems makes analyzing their ethical implications necessary during the design phases of these devices to ensure not only sociotechnical resilience, but to design them for human values in mind and thus preserve meaningful human control over them. This paper engages in a conceptual investigation of how meaningful human control over smart home devices can be attained through design. The value sensitive design (VSD) approach is proposed as a way of attaining this level of control. In the proposed framework, values are identified and defined, stakeholder groups are investigated and brought into the design process and the technical constraints of the technologies in question are considered. The paper concludes with some initial examples that illustrate a more adoptable way forward for both ethicists and engineers of smart home devices.

1. Introduction

In the weeks following his purchase of the *Ring* internet-connected security camera (an Amazon subsidiary) in July of 2019 Alabama man John Baker Orange became the victim of a strange cybersecurity breach, one that unfortunately is becoming ever more common (Noor, 2019). Following his lawsuit against the company, it was revealed that the incident began when a strange voice was heard coming through the microphone of the doorbell camera commenting on Mr. Orange's children who were playing basketball in front of the house at the time (Paul, 2019). The Lawsuit claimed that "unfortunately, Ring did not fulfill its core promise of providing privacy and security for its customers as its camera systems are fatally flawed," noting that the company did not implement two-factor authentication nor put requirements in place to

[†] Institute for Ethics and Emerging Technologies; University of Turin (Consorzio FINO), Italy.

HUMANA.MENTE *Journal of Philosophical Studies,* 2020, 13, 37, 40–65 ISSN: 1972-1293

3 Meaningful Human Control – Two Approaches

John: Jesus, you were gonna kill that guy.

The Terminator: Of course; I'm a Terminator

John: ...You just can't go around killing people.

Terminator: Why?

— Terminator 2: Judgment Day

3.1 Operational Level of Control

Ekelhof's (2019) approach to MHC is predicated on military operational practice, which both supports and constrains targets in areas of operations. This method, though it views MHC as a function of the role of designers, similar to Santoni de Sio et al., and also of technical targeting procedure, as suggested by Leveringhaus (2016), differs in its level of abstraction. It focuses on the higher level of organisation and operational control exercised by the military as a *supra*individual agent. This approach entails that these operational parameters necessarily constrain the 'autonomy' of any AWS (and this too goes for any human agent in the military, such as soldiers). The result is that 'full' autonomy—as is often construed in discussions on AWS—is not 'full' in the sense that is often implied (e.g., self-determining agents), but rather is restricted to various operational decisions and *a priori* planning for deployment and operations.

Ekelhof looks to the case of conventional air operations in order to frame human involvement in operations through a dynamic targeting process. By framing the role of human agent decision-making within distributed systems, he outlines ways in which policymakers and theorists can determine how military planning and operations actually function, and, thus, frame the use of AWS within those practices. In his characterisation of the human role in military decision-making, he unpacks a six-part briefing package (pre-operation), which is thereafter followed by a six-step landscape for mission execution. I briefly summarise these below.

3.1.1 Pre-Mission

The Briefing

Before the mission is undertaken, the air component is briefed with information on mission execution, which can either be highly detailed, including information such as "target location, times, and

munitions", or less detailed, for example when we consider dynamic targeting *in situ* (Ekelhof, 2019, 345). This information is distributed to the various domains of the operation and to specialists, who then vet and use it in order to engage in more detailed planning. The executers of the mission (in this case, fighter pilots) are then brought in, briefed on the mission details, and take the time to study the information provided, while also making any necessary, last-minute preparations for execution. In this briefing package, Ekelhof outlines the following six components that can be included:

1. A description of the target – a military compound – consisting of all available knowledge
2. A target's coordinates
3. A collateral damage estimation (CDE) to provide the operator with an estimation (not certainty) of the expected collateral damage (NATO, 2016). In this example, the risk of collateral damage is low provided the predetermined mitigating techniques are applied
4. A recommendation of the quantity, type, and mix of lethal and nonlethal weapons needed to achieve the desired effects (i.e., weaponeering solution) (USAF, 2017). In our example, these are GPS guided munitions
5. The joint desired impact used as a standard to identify aim points
6. The weather forecast that, in this case, describes a night with overcast condition (clouds cover either most or all of the sky) and heavy rainfall.
   (Ekelhof, 2019, 345).

Coupled with other information—such as the rules of engagement—the operator can then leave to execute the mission.

3.1.2 *In Situ* Operations

Step 1: Find

Intelligence and data are required in order to successfully identify the target for an operation. In this case, such a target is preprogrammed into the fighter jet's navigation system as well as into the payload's navigational system. Whereas a dynamic target requires *in situ* data collection, here the task involves arriving at the preprogrammed "weapon's envelope (i.e., the area within which the weapon is capable of effectively reaching the target)". This process is displayed on the operation's HUD (Ekelhof, 2019, 345).

Step 2: Fix

It is at this stage, once the operator has arrived within the weapon's envelope, that the onboard systems will aim to positively identify the target, which was confirmed during operational planning, in order to ensure that payload delivery is compliant with the relevant military and legal protocols (cf. NATO,

2016). Given that, in this case, the targets were preplanned and confirmed, the operator does not usually engage in visual confirmation for positive target identification. Instead, they rely upon the onboard systems and the validation process that took place during operational planning to ensure that the identified target is lawfully engaged. Therefore, even in this fixed case of preplanning, the human pilot is not required to attend to anything else during this phase of the mission other than arriving within the weapon's envelope (Ekelhof 2019, 345-346).

Step 3: Track

The operator tracks the target within the weapon's envelope to ensure the continuity of positive identification and to provide concurrent updates as to the position/status of the target. In the case of a static target (e.g., a military compound in Ekelhof's example), tracking is relatively straightforward and involves, like in the fix phase outlined above, simply entering into the weapon's envelope (Ekelhof 2019, 346).

Step 4: Target

During this phase, the relevant rules of engagement (RoE), laws of armed conflict (LoAC), and other relevant targeting rules are invoked to ensure lawful targeting and deployment. In addition, other factors are taken into consideration, such as issues relating to collateral damage and risk factors posed to one's own forces. Once again, in this predetermined and validated target case, where the target has already been vetted by legal and military experts, the pilot is permitted to simply input the relevant data into both the vehicle and weapons payload delivery systems to ensure proper execution. In this case, on account of the visually impairing weather conditions, no further collateral damage estimates can be provided owing to the fact that visual confirmation is not able to be made (even if actively sought). Given that the planning at the pre-mission stage had confirmed that collateral damage estimates were low, and that this validation was made in-line with the standard protocols that govern such decisions, the human pilot does not actively participate or intervene in the mission process beyond piloting the vehicle into the weapon's envelope (Ekelhof 2019, 346).

Step 5: Engage

At this stage, once the operator enters the designated weapon's envelope, the onboard computer, based on its knowledge of the equipped weapons system's capabilities, suggests to the pilot the most opportune time to release the payload in order to ensure its effectiveness. Given that the payload

system is GPS guided, there is no need for any other forms of targeting based on visual identification. Once weapon release has been authorised by the pilot, the munitions guide themselves to the target.

Step 6: Assess

At this point, the task is to assess the damage that resulted from the previous stage and to determine the effects of the strike. Naturally, a pilot's visual assessment may be impaired by various factors, such as, in this case, the weather conditions. Likewise, visual assessments of collateral damage from a pilot's vantage point may not accurately reflect the efficacy of the strike and its consequences. In the case of aerial engagements such as this, ground support forces may be required to allow for a more accurate assessment of the engagement (Ekelhof 2019, 346).

3.1.3 Operational Control

When considering MHC, then, it appears that most (if not all) of the performance elements related to each step of the above process are beyond the pilot's control, which could be argued to be emblematic of contemporary aerial operations in general. Whilst the pilot can be said to be in direct operational control of certain aspects of the operation, such as piloting the craft to the weapon's envelope and initiating the weapons release, this type of control is arguably not 'meaningful', in any sufficient sense, given the pilot's potential lack of 'cognitive clarity and awareness' of the situation in which they are participating (Article 36, 2015). This begs the question, then, whether or not the pilot actually does possess sufficient levels of such clarity and awareness as to be deemed substantial in any meaningful way.

Even though discussions at the pilot level may provide some further insight into both operations and modern aircraft that employ AWS, they tend to focus on the wrong vector (i.e., the operator) rather than emphasising how the military, as a supraindividual agent (i.e., an organisation), can maintain MHC over targeting operations. Because of this, the ongoing international debate on AWS tends to overly concentrate on the deployment stage of AWS and their relationship to the individual operators, thereby attempting to locate the vector for MHC between those two agents (AWS-human). In doing so, they ignore the broader covariance in the distribution of labour between agents within the military-industrial complex that make up the decision-making organ. The steps outlined above, and particularly the pre-mission briefing stage with its collateral damage and proportionality assessments, are largely sidelined in these discussions.

What this approach entails, then, is that a distributed notion of agency in MHC is needed to accurately account for the numerous decisions and measures that the different agents in the broader decision-making mechanism undertake prior to deployment. Accordingly, different agents will have different levels of control over any given vector in the process, and any sufficient conception of MHC must reflect this. This, of course, does not negate the role that human operators play, but rather stresses that they form only a part of the larger decision-making network. In this sense, 'full autonomy' is not full in the commonly understood sense but is instead constrained by the larger apparatus of which it forms a part.[15]

3.2 Design Level of Control[16]

The second level of abstraction is drawn from the account of MHC by Santoni di Sio et al (Mecacci and de Sio 2020; Santoni di Sio and van den Hoven 2018). Their account differs from existing approaches to describing MHC by instead providing a philosophical account of MHC, defining it as a covariance between the system's behavior and an agent's decisional intentions and reasons to act. This entails that systems can be designed in a way that permits agents to forfeit some of their direct operational control while still retaining global control of the system. This means that greater, not reduced, levels of autonomy (in certain cases) may actually permit more comprehensive control of a system. As mentioned in the preceding section, more direct operational control does not necessarily constitute being 'meaningful' in the sense that is generally desired with regard to autonomous systems. Attaining MHC in their approach allows for clearer lines of accountability to be drawn when humans remain 'in-the-loop' in relation to these systems, given the fact that tracking the relevant reasons behind an agent's decisions is a necessary condition for MHC.

Their approach to MHC is functionally comprehensive in its scope, looking not only at individual systems but rather at the whole sociotechnical infrastructure of which these systems form a part. This means that although the specific design and deployment of systems have been implicated as important factors in understanding MHC, they cannot be understood in isolation from the infrastructures, organisations, and other agents that are inextricably connected to their design, deployment, and use (Umbrello 2020). The approach is the *design level* because it describes how a system can be purposefully engineered to facilitate MHC. In other words, MHC becomes a technical

[15] This echoes, and Ekelhof repeats it as well, the Defence Science Board's statement that "there are no fully autonomous systems just as there are no fully autonomous soldiers, sailors, airmen or Marines" (USSB, 2012, 23).
[16] Much of the description provided in this section is adapted from a paper I previously published that similarly recounts the account of MHC given by Santoni di Sio et al (Umbrello, 2020).

design requirement, not only of the system itself but also for the relevant sociotechnical infrastructures as well. To do this, however, they outline two necessary conditions that must be met: the *tracking* and *tracing* conditions. Satisfying these two conditions, they argue, permits a more comprehensive conception of MHC to take shape, which reaches beyond solely end users and extends to agents, such as designers and policymakers, as well as organisations, and sets a level of meaningful control and thus clearer lines for attributing responsibility.

3.2.1 Tracking and Tracing Conditions

The tracking condition deals with how responsive a system is to certain actions that are a consequence of human reasoning.[17] It is more comprehensively defined as:

> *First necessary condition of meaningful human control*. In order to be under meaningful human control, a decision-making system should demonstrably and verifiably be *responsive* to the *human* moral reasons relevant in the circumstances—no matter how many system levels, models, software, or devices of whatever nature separate a human being from the ultimate effects in the world, some of which may be lethal. That is, decision-making systems should *track* (relevant) human moral reasons. (Santoni de Sio & van den Hoven, 2018, p. 7)

The tracing condition is different given that it asks if it is possible to delimit the human agent(s) involved in the system's design and deployment history (e.g., designers, manufacturers, users, etc.), who are capable of: (1) understanding the system's potential and (2) can recognise their moral responsibility in relation to a system's deployment and use (i.e., liability of moral consequence). Santoni de Sio and van den Hoven more thoroughly define tracing as:

> *Second necessary condition of meaningful human control*: in order for a system to be under meaningful human control, its actions/states should be traceable to a proper moral understanding on the part of one or more relevant human persons who design or interact with the system, meaning that there is at least one human agent in the design history or use context involved in designing, programming, operating and deploying the autonomous system who (a) understands or is in the position to understand the capabilities of the system and the possible effects in the world of its use; (b) understands or is in the position to understand that others may have legitimate moral reactions toward them because of how the system affects the world and the role they occupy. (Santoni de Sio & van den Hoven, 2018, p. 9)

[17] The use of the term 'reasons' here is understood as any element that can both prompt and demonstrate human behavior, such as objectives, programs, and strategies.

MHC, then, is attained by agents who can satisfy both of these conditions. Only then can they be said to have MHC over a system. AWS, then, can *prima facie* be under MHC by an agent (or agents) if they are designed to support as much as possible the values of accessibility and explicability (explainability and transparency) as manifested in the system's behaviours. If a system is capable of explaining its internal decision-making process (explicability), and such systems are themselves transparent (also a factor of explicability), then such a system can, at least in theory, be more easily brought under MHC given that an agent's (or agents') understanding of the system's use and deployment can be more easily attributed to the system's design architecture.

With these two necessary conditions, MHC ultimately entails a definition of control that is more nuanced and more stringent than operational control, where full direct control is demanded. What makes it more stringent than direct control is that it precludes the attribution of human control to any system merely because it has an agent 'in-the-loop' (e.g., a soldier co-commanding a field operation with an AWS). A commander of an AWS, even if they have a kill switch, or can visibly see the AWS's current status and actions, is not necessarily equipped to understand why the system does what it does. In such cases, MHC by the end user cannot be attained because the tracing condition would not be fulfilled on account of a system's opacity. Although it is true that other agents (e.g., designers, programmers, and/or the state's military institution(s)) may very well understand what is going on in the 'black box' (though this is not always the case). If the system successfully tracks these agents' reasons, and they are deemed to be responsible for and capable of understanding the behavior that the system exhibits based on this tracking, and also for the way it acts based on its tracking of more proximal reasons (as discussed below), responsibility can be attributed to these agents. In other words, they can be said to have had MHC. It is here that we can begin to see how the design level can help to navigate the distributed nature of military operations planning, which has been previously discussed in relation to the operational level of MHC.

Conceptualising MHC in this way is more comprehensive than that of direct operational control for it permits (though it is not a necessary condition) the inclusion of supervisory control, which sanctions the user to supervise a (semi-)autonomous system that is in operational control, yet still permits an end user to intervene in its operation if necessary. Likewise, as already mentioned, this form of direct supervisory control is not a necessary condition for MHC to be deemed to have been attained. A fully AWS can, in principle, be precise, comprehensive, and transparent in tracking the reasons

behind a human agent's decisions in lieu of the ability for human agents to intervene in its operations, thereby still meeting the conditions for MHC.[18]

3.2.2 Distal and Proximal Reasoning

Adopted from the philosophy of intent and action (Bratman, 1984; Mele and William, 1992), Santoni de Sio's and van den Hoven's conception of an agent's (or agents') reasons is further developed, helping to not only specify different types of reasoning within complex systems but also to better understand the inner workings of the tracking condition (Calvert et al., 2018). Calvert et al. (2018) began by developing two distinct types of reasoning: *distal* and *proximal*. Proximal reasons are those intentions that are associated with an action in a temporally immediate way (concurrent), such as the intention to fire upon a target, to stop an imminent strike, or to immediately return to base. Distal reasons are longer term intentions or objectives that are formulated in a less immediate way. A user's distal reason, for example, to use an AWS is to reduce the risk for human operators when engaging enemy combatants, and/or to reduce the economic cost of such engagements. Whereas a company's or programmer's distal reasons may be for the system to adhere to certain contractual norms or to comply with national/international laws (i.e., not permitting an AWS to fire upon surrendering combatants).

	Distal Reasons (longer term, general objective)	**Proximal Reasons (concurrent intentions)**

FULLY AUTONOMOUS WEAPONS SYSTEM	• Plan to Maximise efficiency
	  ○ Reduce briefing-to-deployment time
	  ○ Increase deployment frequency
	  ○ Maximise target accuracy
	• Reduce human error
	• Plan to adhere to IHR Law, Law of Armed Conflict
		• Impromptu intention to have the system return to base
		• Intention to belay a strike given new intelligence
		• Intention to modify payload selection and delivery
		• Intention to change a system's weapons envelope

Table 3. Example of distal and proximal reasons with regard to autonomous weapons systems

[18] What this does, then, is shift the canonical notions of accountability as being a function of the end user to other relevant moral agents within the design history and use of the system.

Distal reasons are those overarching intentions that the relevant agent(s) will have for the desired operations of a system. The concept of direct operational control is naturally aligned and sensitive to proximal reasons, in which a system functions as a consequence of the immediate, concurrent intentions of the human agent. In most cases, these will be the end users who are in proximity to the use of the system. With a (semi-)autonomous predator drone, for example, if the pilot (user) does not release the weapon payload it is because they had no intention, in that instant, to do so (i.e., they could have been distracted or preoccupied with some other task). Because traditional systems like these are – to the best extent possible – under the influence of their human users' proximal reasoning, then those users are causally responsible for their use and consequent impacts. It is for this reason that MHC extends its scope of reasons, to which it must be sensitive to, in order to sufficiently satisfy the tracking condition, particularly in the case of autonomous systems. (Fully) AWS, which we can imagine being connected to various other autonomous systems, such as the information and communication systems of the forward operating base (FOB), the unattended ground systems, and the intelligence, surveillance, and reconnaissance systems, must be sensitive to both distal and proximal reasoning. Satisfying only proximal reasons (i.e., to release the payload or to return to FOB) can come at the cost of more general and objective distal reasons (i.e., reducing friendly casualties).

Mecacci and Santoni de Sio (2020) moved even further beyond this theoretical construct of MHC in order to operationalise it by exploring more concrete design requirements. Taking the work above on the more specific distal and proximal reasons of the tracking condition, they frame MHC as *reason-responsiveness*. It is here that Mecacci and Santoni de Sio make a strong case for the sociotechnicity of autonomous systems given that they broaden MHC as being contingent not only on technical design engineering and more rudimentary human vectors in engineering, but also on the crucial role of institutional design (discussed in greater detail in Chapter 5). Regarding reason-responsiveness more particularly, the complexity of this system (system-within-a-system) refers to the proximity or distance of the various types of human reasoning to the systems' behaviours (Mecacci and Santoni de Sio, 2020). They model complexity of these relations to system behaviors in Figure 1.

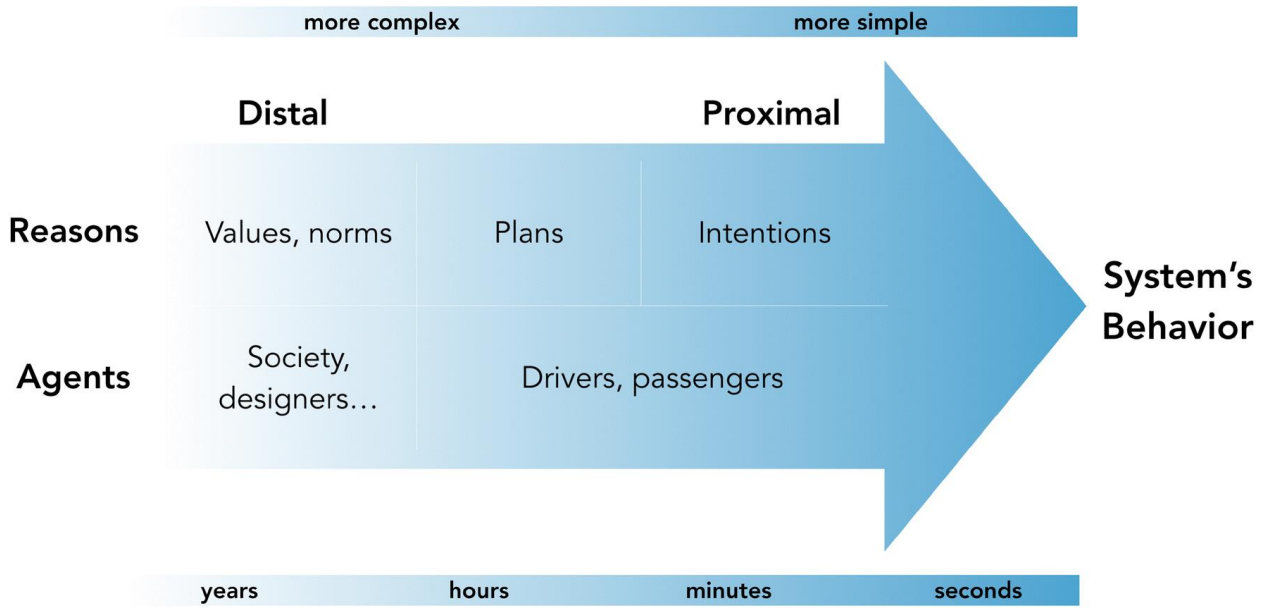

Figure 1. The proximity scale. (Source: Mecacci and Santoni de Sio, 2020)

The above figure is meant to illustrate the relationship of both agents and reasons across times and as functions of complexity. This type of classification is pertinent given that the continuum allows us to more saliently pinpoint the relevant reasons of the relevant agents in any given context. Mecacci and Santoni de Sio point out an important bar here, and that is the temporal factor as it pertains to the reason-responsiveness of systems are markedly different for those of more traditional models of the time dilatation , or lack-there-of, between such intentions and *human* actions (which could be a priori or instantaneous depending on the cognitive study model employed). This model, and as shown in Figure 2, is for *human* intentions and the behaviour of *autonomous systems*. This is made most manifest by the time dilation that can occur between human intention and system action with regards to proximal intentions (as a function of special distance and system lag for example).

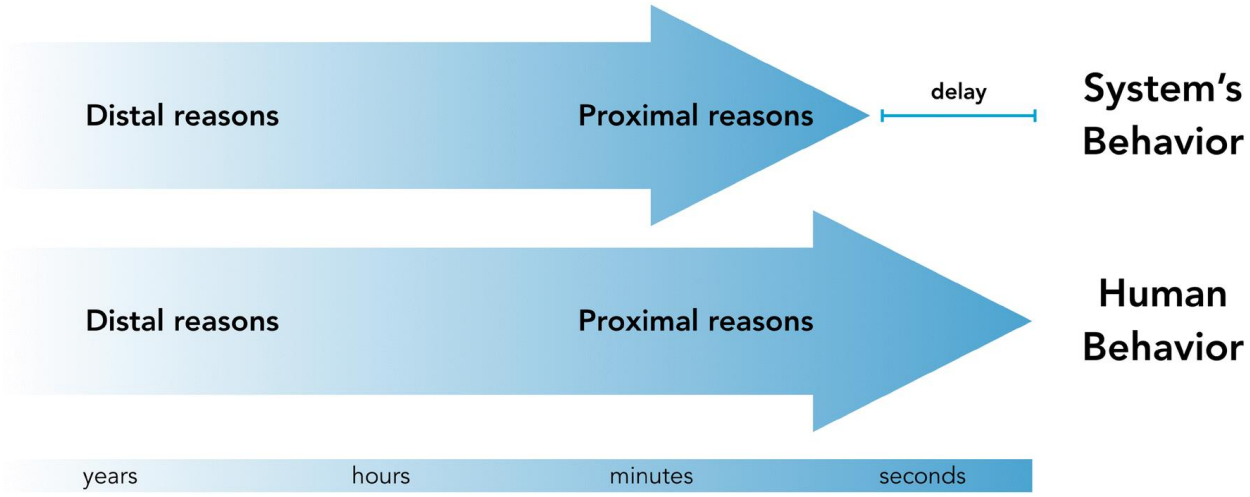

Figure 2. Between human reasons and systems' behaviour there can be a temporal gap which does not compromise the scale. (Mecacci and Santoni de Sio, 2020, p. 110)

The lag, of course, would make itself mostly manifest in terms of a systems' response to proximal reasons given that they are the more specific and temporally immediate reasons. As mentioned above, proximal reasons for a fully AWS may be for the platform to belay an immanent strike whereas the distal reasons explain the systems more general action plan such as entering into the weapons envelope for an aerial strike. This more general reason can of course be decomposed into smaller, more proximal reasons, as a function of the briefing information such as flying at a particular altitude to avoid anti-air missiles or to ensure that changing weather factors do not interfere with onboard navigation and targeting systems. Of course, the more general distal reasons need not, and in many cases will not, be decomposed as such, thus the expression of many potential proximal reasons may not ultimately be articulated in any given operation. This highlights an important point, the above scale allows us to determine the different agents who's reasons are *actually* articulated *if*/*when* they are articulated, and concomitantly, how responsive the system is to those reason (Mecacci and Santoni de Sio, 2020, p. 110). In many cases the proximal reasons like those in Table 1 will be articulated by more direct stakeholders like field commanders whose proximity is smaller in scale. Distal reasons rather may come in the form of superordinate norms from states, treaties, the Laws of Armed Conflict, International Humanitarian Laws, that support and constrain certain operational possibilities.

Designing, however, for the more general and abstract distal reason like those in table 1 are unquestionably more complex in terms of how to design *for* them (i.e., like not causing disproportionate collateral damage). This does not mean that such AWS cannot be sufficiently responsive to distal reasons categorically, in fact current (semi-) AWS already do. Semi-autonomous

drones can already take-off, land, navigate, and travel without human operational control effectively (e.g., General Atomics MQ-9 Reaper drone). However, and this is the philosophically important point here, for an AWS to be meaningfully responsive to the distal reasons, just as it would be to the more specific and technically (relatively) simpler proximal ones, requires that "better automation" (Mecacci and Santoni de Sio, 2020, p. 112). This brings us back to the beginning of this thesis where I propose that more (better) automation, rather than the more intuitive direct (human) operational control can augment MHC rather than exclude it.

If such automation is designed *for* greater reason-responsiveness, then such a higher-level of automation means *more* MHC and not less. What this automation means then is that systems are required to "easily track – that is: *recognize, navigate and prioritise* – the numerous reasons and agents that co-occur in every given situation" (Mecacci and Santoni de Sio, 2020, p. 112).

Notwithstanding, this broadened notion of MHC however is rightly criticised as lacking the higher-level governance structures that account for the institutional and design dimensions of control (c.f., Verdiesen et al., 2020). Verdiesen, Santoni de Sio, and Dignum clearly state that this higher level governance structure:

> is the most important level for oversight and needs to be added to the control loop, because accountability requires strong mechanisms in order to oversee, discuss and verify the behaviour of the system to check if its behaviour is aligned with human values and norms. Institutions and oversight mechanisms need to be consciously designed to create a proactive feedback loop that allows actors to account for, learn and reflect on their actions. (Verdiesen et al., 2020, p.13)

This is undoubtedly the case when considering MHC and, as described above, this higher-level governance structure is argued to be satisfied by the operational level of control. Likewise, the design of both the operational level and design levels are argued to be conducive to being operationalised by the VSD approach in the following part. Their relatively potent approach for an oversight framework for AWS does still leave open gaps for the *actual* mechanisms at the governance level, sociotechnical level ,and technical level for *in situ* governance when an AWS is deployed (see Figure 3). For this reason this project appropriates VSD as the means for framing this central column and designing *for* it.

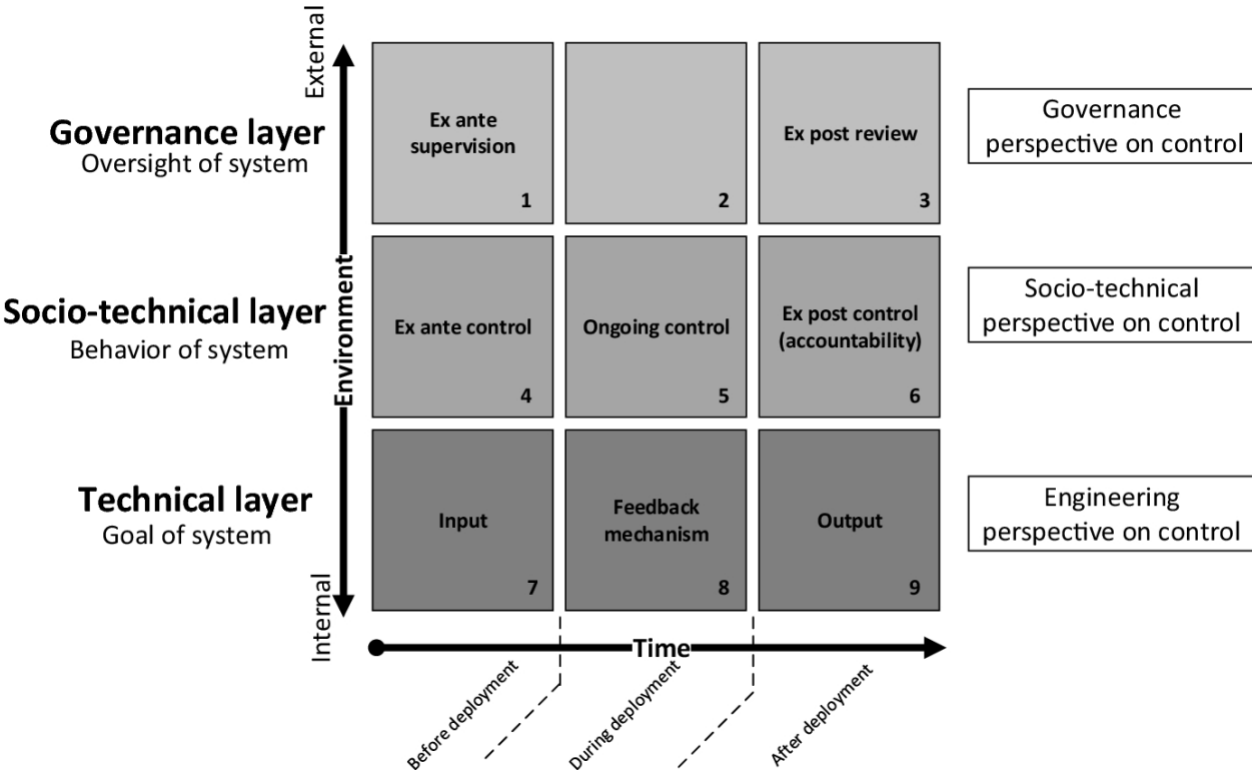

Figure 3. Comprehensive Human Oversight Framework. (Source: Verdiesen et al., (2020, p.18)

Still, adopting such a systems thinking approach to conceptualising the tracking condition requires that all elements that are part of any given system(s) must be maximally sensitive/responsive to the relevant (moral) reasons of any agent, whether they are users or otherwise. This means that it is not solely the burden of agents to be maximally able to behave according to patterns of reasoning, but that every point in a system's infrastructure must be similarly sensitive. This responsiveness can be framed by designers by choosing the proper 'level of abstraction' (Floridi, 2017) in creating autonomous systems (discussed in Part II), which is based on the context of use to ensure receiver-contextualised explanations and transparent purposes (Floridi, Cowls, King, & Taddeo, 2020). This means that any (fully) AWS must not only be responsive to the user's reasons but also conform to established legal and social norms, such as national regulations on the use of autonomous systems, international human rights laws, and the laws of armed conflict amongst others. Mecacci and Santoni de Sio (2020) are explicit in that, although the tracking condition states that the system must be responsive to human reasons and not to other vectors in a system, they argue that social and legal norms reflect the intentions and reasons of supraindividual agents, such as organisations, companies, and states (Mecacci & de Sio, 2020, p. 109). In this case, the operational level of control serves as this supraindividual vector.

3.3 Conclusions

The implications of Santoni de Sio et alia approach are not insignificant, as they appear to run contrary to the notion that greater autonomy entails less MHC. The systems that form the network, which constitutes a fully AWS, and the systems that their integrations subsequently form require comprehensive and ubiquitous design that permits them to be maximally sensitive not only to the end user's intentions and reasons for action, but also to societal norms as well as legal and policy statutes. As already stipulated, such a requirement means having a more stringent notion of what constitutes MHC; however, as a consequence, it permits increased levels of autonomy (i.e., in the case of an AWS, removing human pilots from both physical and psychological harm) with increased control over the system through design decisions as well as operational and regulatory infrastructures. This means that MHC *can* be achieved if systems are maximally responsive to the intentions of agents beyond simply the final users, such as the designers, relevant industries, and states in general (i.e., the military-industrial complex [MIC]).

Despite the nuance in this particular approach to conceptualising MHC (cf. Mecacci and Santoni de Sio, 2020), this dissertation aims to take a more meta-normative approach by combining these theories to produce a more unified notion of MHC for fully AWS. The following chapter begins by discussing how the two LoA are complimentary, how both are underpinned by a systems thinking perspective, and how they can each be optimised via a systems engineering approach via VSD to both operational and design innovation.

References

Article 36. 2015. "Killing by Machine: Key Issues for Understanding Meaningful Human Control." 2015. http://www.article36.org/weapons/autonomous-weapons/killing-by-machine-key-issues-for-understanding-meaningful-human-control/.

Bratman, M. (1984). Two faces of intention. *The Philosophical Review*, *93*(3), 375–405.

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

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

Floridi, L. (2017). The logic of design as a conceptual logic of information. *Minds and Machines*, *27*(3), 495–519.

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). Designing AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*, 1–26. https://doi.org/10.1007/s11948-020-00213-5

Leveringhaus, Alex. 2016. "Drones, Automated Targeting, and Moral Responsibility." In *Drones and Responsibility: Legal, Philosophical, and Socio-Technical Perspectives on the Use of Remotely Controlled Weapons*, edited by Ezio Di Nucci and Filippo Santoni de Sio, 169–81. Routledge. https://doi.org/9781138390669.

Mecacci, G., & de Sio, F. S. (2020). Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics and Information Technology*, 22:103–115.

Mele, A. R., & William, H. (1992). *Springs of action: Understanding intentional behavior*. Oxford University Press on Demand.

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

NATO STANDARD AJP-3.9 ALLIED JOINT DOCTRINE FOR JOINT TARGETING Edition A Version 1. (2016).

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

USAF. (2017). *Annex 3-60 Targeting*. https://www.doctrine.af.mil/Doctrine-Annexes/Annex-3-60-Targeting/

USSB. 2012. "Defense Science Board Task Force Report: The Role of Autonomy in DoD Systems." Washington, DC. https://doi.org/ADA566864.

Verdiesen, Ilse, Filippo Santoni de Sio, and Virginia Dignum. 2020. "Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight." *Minds and Machines*. https://doi.org/10.1007/s11023-020-09532-9.

4 Coupling Levels of Abstraction - A Two-tiered Approach

4.1 Technical Full Autonomy and AWS

As mentioned in the introduction, one of the central premises on which proponents of a ban on AWS base their case relates to the concern that certain increased levels of autonomy may result in an accountability gap in the event of recalcitrance. Sharkey (2014) aptly describes five levels of technical autonomy that can describe AWS targeting (Figure 1). The least problematic stage is Level 1 (although Ekelhof's (2019) analysis arguably brings into question 'which' human). Levels 4 and 5 are argued to be the most problematic. Level 4, like Level 5, is argued to be dangerous given 'how' an AWS selects a target (i.e., systemic opacity, computer vision, etc.) and its technical ability to do so as a function of the various targeting norms and rules of engagement. Similarly, the fourth level brings into question the cognitive clarity of the human operator, who has veto power and the ability to determine the validity of the system's chosen target(s). Regardless, Level 5 is typically the subject of debate as it is considered the key descriptor of full autonomy in terms of AWS.

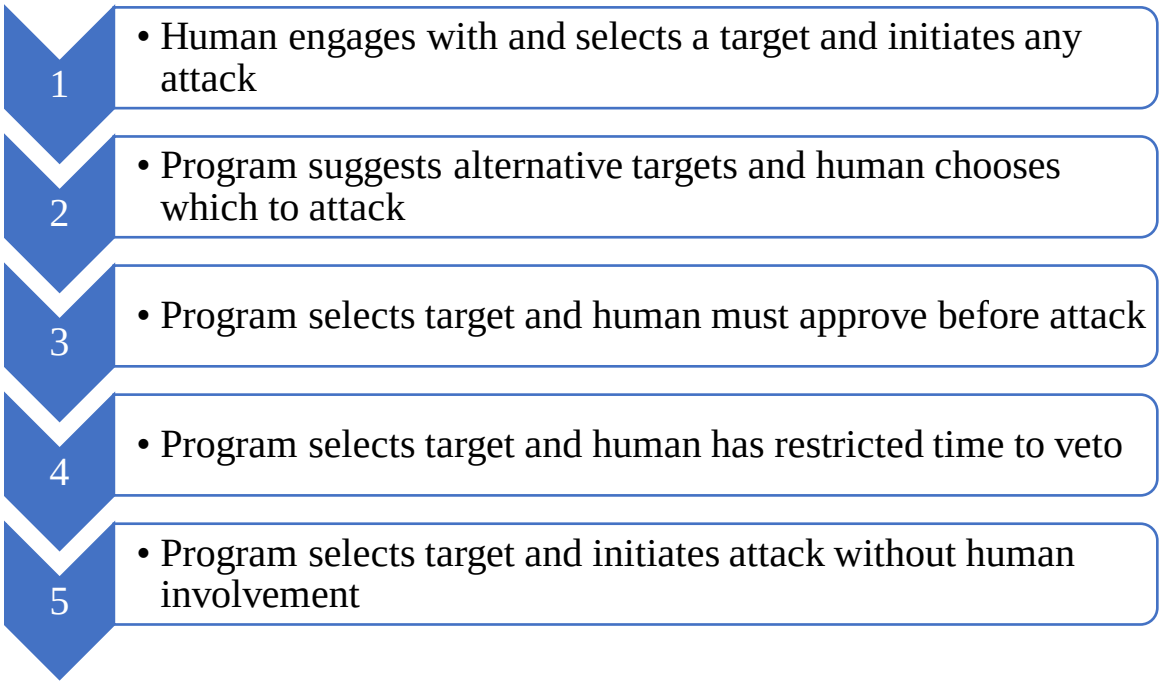

Figure 1. Level of Autonomy. Source: (Sharkey, 2014).

Here we can already begin to tease out some of the potential issues that exist with problematising autonomy. Though there are convincing arguments against AWS, other than the supposed accountability gap proposed by the above ordinance, such as the dehumanisation of war and its deleterious effects on human dignity, or even the functional necessity of lethality, it appears that

actual military operations planning and deployment strategies intuitively constrain the autonomy of any given agent, soldier, or AWS, so as to be a function of a larger *a priori* plan that bears little, if any, intrinsic operational value outside the functional capacity to be able to carry out such plans. This, of course, does not preclude AWS deployed within such constraints from limitless actions or from wanton recalcitrance. The technical design, which is predicated on the technical design requirements, must reflect both the proximal and distal intentions (i.e., reason-responsiveness) and goals of the relevant agents within the deployment envelope. These would be the commanders who employ such weapons in their area of operations, as well as the potential human operators who may be engaging with them on the ground (i.e., they can be aerial AWS, e.g., fully autonomous drones/fighters). Regardless, the capacity for these systems to be responsive to the relevant moral reasoning of the agents involved must be considered as a foundational variable in the weaponeering decision-making process for any given context of deployment in the pre-mission stages. And it is institutional processes like weaponeering that *de facto* predicate a level of *a priori* operational control like that suggested not only by Ekelhof (2019), but also by Verdiesen et al. (2020) as part of the 'before deployment layer (c.f., Chapter 3, Figure 3).

4.2 Coupling Levels of Abstraction for MHC

In practice, then, systems thinking provides salient grounds for thinking about these various LoA. The procedural process of operational planning and target identification form the higher (or meta-) level of MHC, as clearer lines of causality can be conceptualised, culminating in weapons release and efficacy assessments. This level, of course, can be further broken down into more granular LoA like *strategic, tactical*, and *operational*, but those would just be more compartmentalised categories for the umbrella of military operations. Similarly, the design level of MHC is functionally dependent on a system's understanding of both *tracing* design histories as well as *tracking* the responsiveness of autonomous systems to the relevant moral reason(s) of the relevant agent(s) in the design and use chains of such systems. Theoretically speaking, both LoA are predicated on systems or networks of interconnected nodes (Figure 2). Similarly, both LoA, despite their different scopes, feed into one another. Within the operational level, the bounds within which weaponeering decisions are made prior to deployment are contingent on the functionality of the system itself, in order for it to be chosen as the most appropriate means for carrying out the intended mission. However, such technical responsiveness to on-the-ground needs for successful mission completion is not contingent on those types of pre-mission assessments. System-level recalcitrance can jeopardise the overall level of MHC despite the system being bound by

the operational level of control. For this reason, weaponeering decisions must be reflected in the design level in order for those decisions to be sufficiently salient prior to deployment. Thus, the operational level feeds down into the design level by supplying the norms, objectives, and intentions necessary for deployment to be lawful, and for the operational level itself to be holistic in terms of retaining sufficient control (this is illustrated in Part II). Likewise, the various agents who are essential to the pre-mission planning stage of operations form *part* of the number of relevant moral agents (or, collectively, of the supraindividual agent) that permits the design level to *actually* design AWS to be sufficiently responsive to the reasons and intentions of those actor(s), which makes the weaponeering of AWS permissible and, thus, under *a priori* MHC on both LoA. This would, of course, mean tightening existing military-industrial partnerships that use these agents as stakeholders for whom these systems can be designed *for* (coupled with the relevant RoE and LoAC). This seems intuitively necessary to increase the equifinality of the complex relations of the different agents that form these collaborating institutions and more saliently track the reasons within such networks of systems-within-systems.

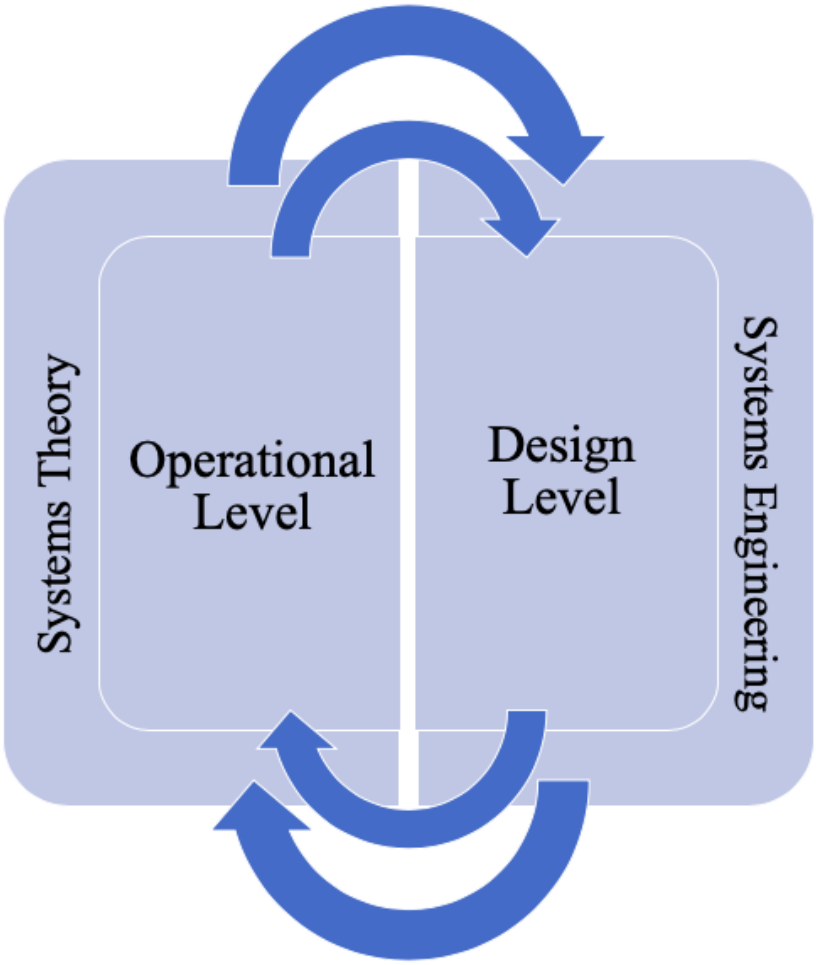

Figure 2. Superordinance of Systems Theory and Engineering over the two levels of abstraction of MHC.

If we envision, then, a scenario that is often discussed in the literature against AWS – the case of an AWS killing civilians – we can begin to trace reasons for dismissing this *prima facie* objection. If an AWS kills a civilian on the ground, who is within the weapons envelope that was delimited prior to deployment, the killing is not *mala in se* to the extent that collateral damage assessments have already been agreed upon during the pre-deployment stage and under the existing norms regarding proportionality. To some extent, the killing of civilians is not necessarily equivocal to recalcitrance and can instead be traced back to the briefing information (see Chapter 5 on the *wrong reasons problem* of *actualised* values). If we imagine that, even within the weapons envelope, an AWS kills civilians disproportionately, and over and above the acceptable level of collateral damage determined during pre-planning, this can be construed as technical recalcitrance, which can be traced back to the relevant agents within the design and use histories of the AWS to determine if the system was designed in such a way so as to be maximally responsive to the relevant intention(s) of those agents. If this is shown not to be the case, then the AWS cannot be said to be under MHC and is, therefore, not only an unviable option for weaponeering decisions but is also unlawful as well (this is a good vector for thinking about ban criteria). If the relevant agents, such as the designers and users (commanders, AWS designers/programmers, proportionality specialists, MIC as a whole, etc.), are capable of understanding the system's capabilities and the consequences of its use, then they can be said to be in possession of MHC, both in their weaponeering decisions (on the operational level) as well as the design decisions (at the design level). Divorcing one level from the other leaves open vectors from which accountability gaps can arise (c.f., Verdiesen et al., 2020).

Systems engineering, then, can be said to be the design and applications of both of these levels of MHC. Of course, systems engineering seems, and perhaps is, more appropriately spoken of in terms of the design level, given its explicit focus on building autonomous systems responsibly, in a holistic, and anticipatory way, and specifically aligned with the values of the relevant stakeholders. However, it is for this reason that the operational level is necessary, and, as mentioned above, the complex network of agents, who are distributed across the military target acquisition and deployment process, are all relevant moral agents making up part of the larger system of which the AWS forms a part (i.e., the MIC). Likewise, AWS themselves are systems that are embedded in the larger sociotechnical network of operations and those human agents involved. Analysing and eliciting the needs of these human agents in order for them to make informed and (hopefully) lawful decisions in terms of weaponeering, both in the early stages of and throughout the development cycle, is critical – while always keeping in

mind the whole of the systems' lifecycle, as it relates explicitly to those weaponeering decisions rather than as a discrete technological artifact divorced from its use-context. In other words, rather than building AWS and marketing these systems as novel weapons platforms, both designers and experts, who are involved in the planning of operations, must themselves be part of the design team to weaponeer the design decisions themselves (i.e., part of the population of direct stakeholders in co-creation and co-design in VSD).

If we take this into account when looking at the issues that are often presented by those against the development of AWS, many of the technical issues that are presented as *mala in se*, such as increased autonomy (particularly Level 5, as in Figure 1) or the targeting of civilians, are only problematic if decoupled from a responsible design process and actual military planning and operational practices. When considering these, the augmentation of autonomy is necessarily constrained by many (if not all) of these processes; and, in certain cases, can increase, not decrease, the ability to have MHC. If these systems are designed in such a way as to be maximally sensitive to the relevant moral reasons of the relevant moral agent(s) involved, then they likewise augment MHC, not lessen it. Technical autonomy is often the mechanism by which this the augmentation of this sensitivity can and should be design *for*. Mecacci and Santoni de Sio (2020) aptly demonstrate this seemingly paradoxical paradigm by looking at the example of autonomous vehicles. The marriage of both LoA, then, is teleological as it drives towards systemic synergy in order to avoid component friction and, subsequently, to avoid any unreliability in the design and deployment of AWS.

For systems engineering practices to be successful in optimising equifinality across the various levels of nesting, then complexity has to be modeled as a function not only of the technical architecture of a system (i.e., AWS) but also the logistical human organisation of data (i.e., planning, target data, proportionality assessments, geography, etc.). Because of this, systems can become increasingly complex given the volume and quality of the data, variables, and components across both technical and human spheres. Much of this can be addressed through the design and development of smarter control algorithms and environmental systems analyses; while tools such as system architecture modeling, verification and learning simulations, statistical and reliability analyses, as well as formal decision-making psychology can all be levied to understand the covariance between technical design and human operations. Divorcing the operational level from the design level leaves design impotent and potentially recalcitrant, ignoring most of the *actual* processes that are taken place within this particular context of use. Divorcing the design level from the operational one leaves operations with an opaque and nebulous lethal tool that may result in poor, possibly even unlawful, weaponeering decisions. Thinking

about systems and, more specifically, of these various levels of abstraction as mutually co-constituting one another permits the inherent complexity of these systems to be more easily modeled and consequentially designed *for*, rather than leaving design decisions as *ad hoc* afterthoughts. Doing this allows for clearer lines of emergent behaviors and boundaries to be traced, provided that systems thinking is employed at all levels of nesting.

4.3 Limitations and Specifying the Nexus of MHC for AWS

There is at least one notable limitation of this multi-tiered approach, which is the dynamic engagements of AWS *in situ* rather than in the case of purely pre-programmed engagements (echoed also by Verdiesen et al. (2020) in their framework's central column). This is particularly true of ground-based AWS in comparison to aerial ones. Ground-based AWS can (and likely most often will) find themselves in dynamic and changing engagement scenarios, even within the weapons envelope. Their ability to adhere to the determined mission objectives and targets, while also adapting to an evolving scenario, poses both technical and ethical issues. It appears that such types of ground-based systems take on more agency given the decisions and processes for identification that emerge from dynamic war theatres and their proximity (and thus more fine-grained situational input) to targets. The operational level may be insufficient for grounding MHC in such cases, yet the design level can still provide possible ways of ensuring sufficient control. If the systems are designed in such a way as to be maximally reason-responsive to the largest set of moral reasons and intentions of the relevant agents – perhaps in this case the commanding officer on the ground alongside the AWS and/or the commander supervising the mission/engagement (i.e., continual monitoring) – then recalcitrance of such systems can be tracked and traced back to these individuals, as well as to the designers who originally engineered the autonomous systems (i.e., the military-industrial complex as a supraindividual agent). The real difference here between aerial AWS weapons is that there is epistemic distance between the system and the target, something that is given as a particular variable to proportionality and discrimination assessments. This, of course, if much different for ground-based AWS which have a significantly smaller epistemic gap with regards to target acquisition and engagement. This, in turn, changes how the operational level can be applied given the proximal closeness of a system to its targets that does not necessarily exist between aerial systems and their targets.

Either way, what this limitation at least shows is a further nuance that undermines arguments for a blanket ban on (fully) AWS. That is to say, the difference between aerial (fully) AWS and ground-based (fully) AWS. The operational level seems to tokenise the agent that is in direct

operational control of the engagement and strips them of most (if not all) of the relevant levels of autonomy that are necessary for moral responsibility. For this reason, the substitution of such human agents in aerial engagements appears, at least *prima facie*, benign. For a ban on (fully) AWS to be effective, then, it seems that targeting autonomy *per se* is not the right strategy, at least based upon the above example. Instead, a more effective route would be targeting various specific types of AWS and differentiating between them (i.e., ground, aerial, naval AWS). Of course, this risks over-specifications and leaves open the possibility of circumventing very specific designations and criteria for banned systems. However, this should not discount the above criticism, rather it should encourage wrestling with it head-on in order to ensure more robust policy making.

4.4 Conclusions

Part I of this thesis uses systems thinking and systems engineering as conceptual tools to frame the commonalities between two different levels of abstraction in understanding meaningful human control of autonomous weapons system. It argues that, with AWS in particular, both LoA are necessary for having MHC of AWS. If this coupling is successful then the argument that increased levels of autonomy are problematic, which is at the foundation of most calls for a ban on those types of AWS, is greatly weakened and perhaps even negated entirely. It does this by showing how autonomy, whether human or that of an AWS, is necessarily constrained by military operational planning and the co-construction of these systems with the involvement of relevant moral stakeholders. As long as strict conditions are met across LoA then increasing the autonomy of AWS to what is traditionally called 'full' autonomy is not problematic, and can, conceptually, also increase MHC.

References

Ekelhof, M. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy*, *10*(3), 343-348.

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

Sharkey, N. (2014). Towards a principle for the human supervisory control of robot weapons. *Politica & societa*, *3*(2), 305-324.

PART II

DESIGNING MEANINGFUL HUMAN CONTROL WITH VALUE SENSITIVE DESIGN

AI and Ethics
https://doi.org/10.1007/s43681-021-00038-3

ORIGINAL RESEARCH

Mapping value sensitive design onto AI for social good principles

Steven Umbrello[1] · Ibo van de Poel[2]

Received: 23 October 2020 / Accepted: 30 November 2020

Abstract
Value sensitive design (VSD) is an established method for integrating values into technical design. It has been applied to different technologies and, more recently, to artificial intelligence (AI). We argue that AI poses a number of challenges specific to VSD that require a somewhat modified VSD approach. Machine learning (ML), in particular, poses two challenges. First, humans may not understand how an AI system learns certain things. This requires paying attention to values such as transparency, explicability, and accountability. Second, ML may lead to AI systems adapting in ways that 'disembody' the values embedded in them. To address this, we propose a threefold modified VSD approach: (1) integrating a known set of VSD principles (AI4SG) as design norms from which more specific design requirements can be derived; (2) distinguishing between values that are promoted and respected by the design to ensure outcomes that not only do no harm but also contribute to good, and (3) extending the VSD process to encompass the whole life cycle of an AI technology to monitor unintended value consequences and redesign as needed. We illustrate our VSD for AI approach with an example use case of a SARS-CoV-2 contact tracing app.

Keywords Value sensitive design · VSD · Artificial intelligence · AI4SG · Sustainable development goals · COVID-19

1 Introduction

There is ample discussion of the risks, benefits, and impacts of Artificial Intelligence (AI). Although the exact effects of AI on society are neither clear nor certain, AI is doubtlessly having a profound impact on overall human development and progress and it will continue to do so in the future [1–3]. AI is understood here as a class of technologies that are autonomous, interactive, adaptive, and capable of carrying out human-like tasks [4]. We are particularly interested in AI technologies based on Machine Learning (ML), which allows such technologies to learn on the basis of interaction with (and feedback from) the environment. We argue that the nature of these learning capabilities poses specific challenges for AI design. AI technologies are more likely than not to acquire features that were neither foreseen nor intended by their designers. These features, as well as the ways AI technologies are learning and evolving, maybe opaque to humans [5].

In this article, we build on and extend an approach to ethical design called value sensitive design (VSD). Although other tools for achieving responsible research and innovation have been proposed [6, 7], we specifically chose VSD as the design methodology due to its inherent self-reflexivity. VSD also emphasizes an engagement with both direct and indirect stakeholders as a fundamental part of the design process and the philosophical investigation of values [8, 9].

Past research has explored how VSD can be applied to specific technologies such as energy systems [10, 11], mobile phone usage [12], architecture projects [13], manufacturing [14, 15], and augmented reality systems [16], to name a few. Similarly, it has been proposed as a suitable design framework for technologies emerging in both the near- and long-term future. Examples include the exploratory application of VSD to nanopharmaceuticals [17], molecular manufacturing [18], intelligent agent systems [19], and less futuristic autonomous vehicles [20, 21]. Although these studies provide a useful theoretical basis for

✉ Steven Umbrello
steven.umbrello@unito.it

Ibo van de Poel
I.R.vandepoel@tudelft.nl

[1] Institute for Ethics and Emerging Technologies, University of Turin, Via Sant'Ottavio, 20, 10124 Turin, Italy

[2] Delft University of Technology, Faculty of Technology, Policy and Management, Jaffalaan 5, 2628 BX Delft, The Netherlands

Published online: 01 February 2021

5 Value Sensitive Design: Conceptual Challenges Posed by AI Systems

Progress, not perfection — (Friedman et al., 2017)

5.1 Introduction: A Recap of Value Sensitive Design (VSD)

Wherein lies the problem lies also the solution. *Mutatis mutandis* to the philosophical substructure of the species of AWS discussed in the preceding section (i.e., certain aerial fully AWS), level 5 autonomy (c.f., Sharkey, 2014) becomes non-problematic. It even becomes the key to MHC. When these levels of abstraction for understanding MHC are coupled, aerial fully AWS *de facto* fall under MHC.

One could levy the argument that such a conception of MHC is too philosophically abstract – so abstract that it is rendered incapable of being designed in any way attainable by engineers and MIC partnerships. As discussed, autonomous systems are fundamental to the network-centric warfare (NCW) doctrine that characterises much of what we now understand as modern warfare. These evermore complex systems of people, infrastructures, and organisations (i.e., systems-within-systems) complicate things even further, thus making the problem of many hands more daunting (Taylor, 2020). But what the preceding section aimed to show is how, in principle, the coupling of those two levels of abstraction for understanding MHC allows us to more saliently identify the groups (and thus the individuals within these groups) that can be held responsible for these systems. This bridges the potential responsibility gap of AWS not under MHC.

However, there remains lacunae to traverse. How can we actually design for this particular conception of MHC? The complexity of the design endeavor stems not only from the complexity of the systems that characterise MHC, but also from the unique challenges that AI systems generally pose to the responsible design and innovation of systems employing such AI systems. In both scholarship and popular culture, there is ample discussion of the risks, benefits, and impacts of AI. Although the exact effects of AI on society are neither clear nor certain, AI is and will doubtlessly continue to have a profound impact on the flourishing of humanity (Baum, 2016; Floridi et al., 2018; Winfield et al., 2019). When it comes to AWS, Part I describes the determining system behind what makes fully AWS fully autonomous in the technical sense. Here, AI is understood as a class of technologies that are autonomous, interactive, adaptive, and capable of carrying out human tasks (Floridi & Sanders, 2004). I pay attention to AI technologies based on machine learning (ML) in particular, as the latter allows the former to learn from interaction with, and feedback from, its environment. I argue that these learning

capabilities pose specific challenges for the design of AI. AI technologies are more likely than not to acquire features that were neither foreseen (or even foreseeable) nor intended by their designers. These features, along with the way such technologies learn and evolve, can be opaque to humans (Boscoe, 2019).

In this chapter, I build on and extend an approach to ethical design known as Value Sensitive Design or VSD (see Annex II). There have certainly been proposals for other tools to achieve responsible research and innovation (Initiative, 2016; UNESCO, 2017). But I chose VSD as a design methodology for its inherent self-reflexivity. Furthermore, it emphasises engagement with both direct and indirect stakeholders as a fundamental part of the design process and the philosophical investigation of values (Friedman & Hendry, 2019; Umbrello, 2018). Past research has explored how VSD can be applied to specific technologies such as energy systems (Mok & Hyysalo, 2018; Mouter et al., 2018), mobile phone usage (Woelfer et al., 2011), architecture projects (van den Hoven, 2013), manufacturing (Longo et al., 2020), and augmented reality systems, to name just a few (Friedman & Kahn Jr., 2000). It has similarly been proposed as a suitable design framework for future technologies, both near and long term. Examples include its exploratory application to nanopharmaceuticals (Timmermans et al., 2011), molecular manufacturing (Umbrello, 2019), intelligent agent systems (Umbrello & De Bellis, 2018; van Wynsberghe, 2013), and less futuristic autonomous vehicles (Calvert et al., 2018; Thornton et al., 2018). Although these studies provide a useful theoretical basis for how VSD might be applied to specific technologies, they do not account for the unique ethical and technical issues that various AI systems present.

To address these challenges, both Ibo van de Poel and myself suggest adding a set of AI-specific design principles to VSD (Umbrello & van de Poel, 2021). We propose building on significant headway made recently in numerous AI for Social Good (AI4SG) projects becoming popular in various research circles. Practical, on-the-ground applications of AI4SG principles have already been enacted for various AI-enabled technologies (Mabaso, 2020) This provides researchers with solid groundwork on how ethics can be manifested in practice. But AI4SG is difficult and its underlying principles are still fuzzy, given the multiplicity of research domains, practices, and design programs (Taddeo & Floridi, 2018). Nonetheless, some work has already been done to narrow down the essential AI4SG principles (Floridi et al., 2018, 2020).

5.1.1 Value Sensitive Design

To demonstrate the applicability of the VSD approach to AWS design, it is worth revisiting the approach itself. For the more interested reader, Annex II provides a literature review on the history, issues, and applications of the VSD approach more generally.

VSD is a principled approach to taking values of ethical importance into account in the design of new technologies. The original approach was developed by Batya Friedman and colleagues from the University of Washington. As VSD adoption grew more widespread, it was developed further by others – sometimes under somewhat different headings such as 'Values at Play' or 'Design for Values' (Flanagan et al., 2008; van den Hoven et al., 2015). At the core of the VSD approach is what Friedman et al. (2008) call the tripartite methodology of empirical, conceptual, and technical investigations (see Figure 1). These investigations can be carried out consecutively, in parallel, or iteratively. They involve: 1) empirically investigating relevant stakeholders, their values, and their value understandings and priorities; 2) conceptual investigations into values and possible value trade-offs; and 3) technical investigations into value issues raised by current technology along with the possible implementation of values into new designs.

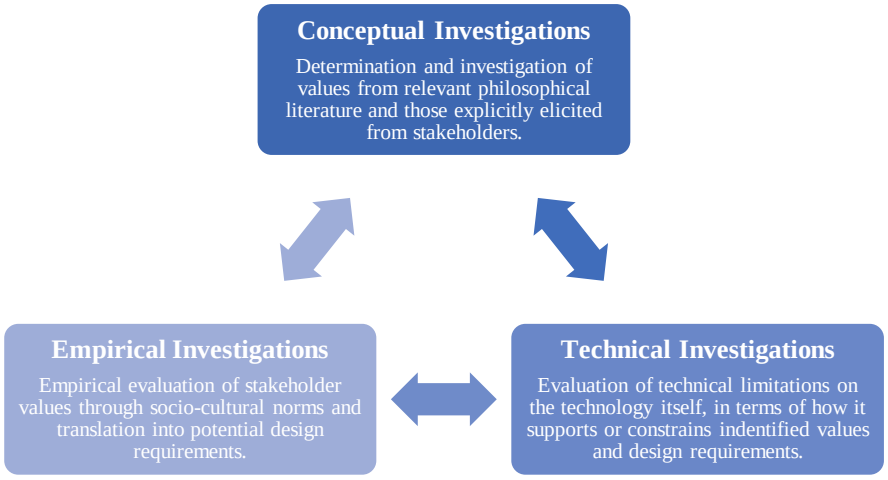

Figure 1. The recursive VSD tripartite framework employed in this study. Source: Umbrello (2020b).

One important issue in VSD is how to identify the values that should be taken into account in a concrete VSD process, as discussed in greater detail in 6.2 (Davis & Nathan, 2014). Friedman et al. (2017) propose a list of thirteen values important to 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 have

opposed such an approach, arguing that it is better to elicit values from stakeholders in a bottom-up fashion (Borning & Muller, 2012; Le Dantec et al., 2009). Both approaches probably have their advantages and disadvantages (c.f., Umbrello, 2020b). For instance, a more general list could overlook values that are important in specific situations. Although bottom-up elicitation can help uncover such values, it is also not a watertight solution as stakeholders may fail to articulate important values (or crucial stakeholders may not have been identified). Moreover, not all values held by stakeholders carry ethical importance that should be included in VSD.

When it comes to identifying values in VSD design processes for AI technologies, some considerations are important. There is now widespread consensus that AI raises specific ethical issues that are not raised (or at least raised to a much lesser degree) by more conventional information and communication technologies (Floridi et al., 2018). This has several implications for the issue of value identification. For one, the original VSD list of values does not suffice for AI. Instead, one could take the values identified by the EU High-Level Expert Group on Artificial Intelligence (HLEG) as starting point (Floridi, 2019; High-Level Expert Group on AI, 2019): *respect for human autonomy*, *prevention of harm*, *fairness*, and *explicability*. For another, some value list would seem desirable in the case of AI simply to ensure that typical ethical concerns arising from AI are not overlooked. This is not to say that no other values should be included in the design of AI applications. Perhaps they should be for specific context. But they should (and some form of bottom-up elicitation may be relevant here[19]) certainly be supplemented by principles to ensure typical ethical issues in AI are properly addressed. This is especially true for highly specific deployment domains, such as the military in this case, where context-specific values need to be accounted for. The proposal, then, is recourse to the AI4SG meanings and factors discussed in Chapter 6.

5.2 Intended, Realised, and Embodied Values of Sociotechnical Systems

Since the empirical and (now) design turns within the applied ethics of technology, the notion that ‘artifacts have politics’ has been a fundamental philosophical precept (Winner, 2003). More clearly, sociotechnical systems *embody* values (van de Poel, 2020; van den Hoven, 2017; van den Hoven et al., 2012). This underlying precept also extends to VSD. AI systems are no exception to such embodiment given that they, too, are sociotechnical systems. But the added complexity of self-learning and typically

[19] Bottom-up approaches can be informed by the actual process of participatory design and responsible research and innovation such as those by (Abebe et al., 2020; Liao & Muller, 2019; Smith & Iversen, 2018; Whitman et al., 2018) as well as the emerging regulation on constraining data collection practices and the design of AI systems e.g., regarding "protected characteristics", human oversight, and informational roles (Smith & Iversen, 2018; Wachter & Mittelstadt, 2019).

opaque inner workings in AI makes the question of a system's value compliance (or recalcitrance) to the various codes of AI ethics (e.g., HLEG, IEEE, etc.) a prescient concern. Ibo van de Poel (2020) categorises three different understandings of value compliance in sociotechnical systems: (1) *intended values* or IV, (2) *realised values* or RV, and (3) *embodied values* or EV.

For the first, the *intended value* of compliance is understood as guiding AI designers in the design of AI systems. Compliance is also integrated into the design of AI systems by those designers so as to align with their intended values as best as possible (van de Poel, 2020). However, there is a privation in this type of compliance; AI systems can still feasibly be compliant in the sense that the intended value may exist in the system, even if it is not fulfilled in any meaningful way.

Consequently, *realised values* focus on the lacuna of (1). Such values understand compliance through a focus on the actual values expressed by the operation of an AI system. But this approach is not without its own issues. Because the focus here is on the actual values expressed by the operation of a system, the system must be deployed first. Only then can adjudication of its value compliance be determined (this is markedly deleterious for warfare systems) . Ideally, such compliance would take place prior to rollout to ensure those values are not detrimental or expressed in deleterious ways. Another issue lies in the immanence of *realised values* in that they are too *prima facie*. Not all system behaviours can be mapped out as a meaningful realised value onto the system itself. For example, imagine an aerial fully AWS naturally employing the AI systems that make it functionally and fully autonomous. The fully AWS is the direct cause of an aerial strike that kills a disproportionate number of civilians relative to the military objective. Would this automatically mean that the AWS was recalcitrant due to the *realised value* of non-compliance (i.e., recalcitrance in relation to the principle of proportionality) obtained? Intuitively, it would seem a case such as this does not warrant such a partisan stance. The failure of the strike in terms of its proportionality is not necessarily attributable to the operation of the system itself. But it may be the cause of other factors, such as the *a priori* intelligence gathered, the proportionality assessment, munitions weaponeered prior to deployment, and/or changing circumstance attributable to mitigating factors on the ground, and so on.

In both of these understandings, the issue is a problem of compliance with the wrong kind of reasons (c.f., Jacobson, 2013; van de Poel, 2020). This problem highlights issues in the loci of the reasons themselves. *Intended values* are problematic because reasons for or against those values are predicated on the intentions of designers (and their subsequent biases, root values, etc.) rather than on the actual AI systems themselves. Similarly, *realised values* are problematic because the reasons may be predicated on improper and/or unprecedented use. Such reasons are therefore also not predicated on

the AI system per se (i.e., the multi-use nature of technology). Hence, the loci of both forms of compliance are located outside the designed system.

To avoid the wrong kind of reasons problem, van de Poel (2020) argues for a focus on *embodied values* (see Figure 2). Embodied values "should be understood as values that have been intentionally, and successfully embedded in an AI system by its designers" (van de Poel, 2020, p. 390). In order to fulfill embodiment, two conditions must be met:

1. Designers must intentionally design AI systems such that they comply with any given or set of values; and
2. When correctly used, the AI system has to *actually* map on to or support the values of (1).

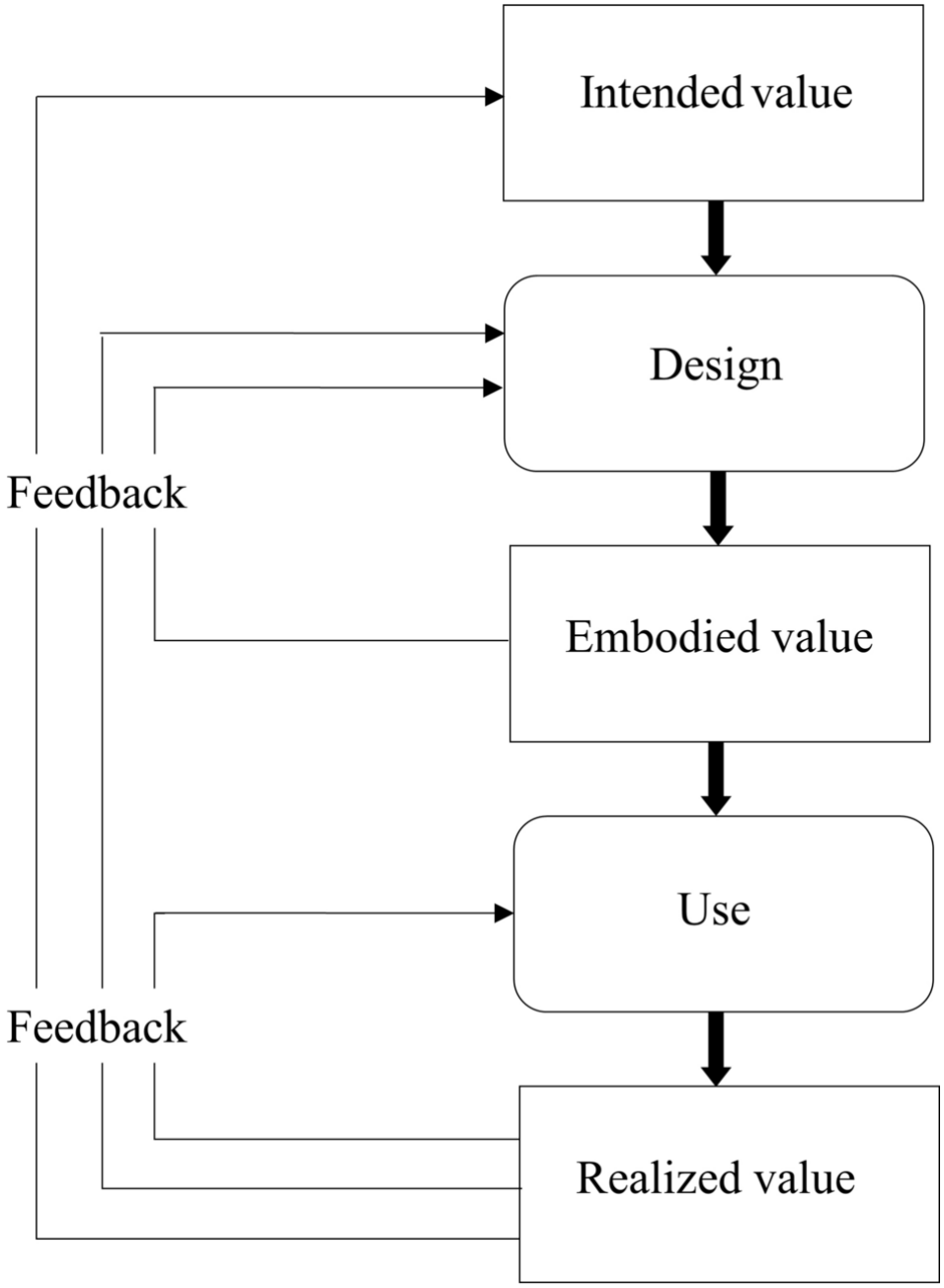

Figure 2. The relation between intended, embodied, and realised values (Source: van de Poel, 2020, n. adapted from Fig. 7.1 in van de Poel and Kroes, 2014).

Intended values alone may ultimately be incongruent with the *actual* operation of an AI system (or any other designed system, such as an institution like the MIC) that expresses different values. These values may also differ – and usually do – as a result of poor design choices. *Embodied values* are the outcome of *intended values* (by the design team) combined with actual *realised values* as a consequence of proper use. The dynamic nature of the conditions that constitute embodied values

(intended + actual) results in continual feedback, as shown in Figure 2. In VSD, this type of change constitutes a trigger for redesign in light of potential issues or value incongruencies (further discussed in Chapter 6). Three kinds of feedback loops may then be triggered, as illustrated in the three levels of directional arrows in Figure 2 above:

1. IV = EV → system use can be modified with the modification of design
2. EV ≠ IV → the modification of design is necessary
3. RV = unforeseen consequences → IV may need to be modified

This dynamism illustrates how the design programs of sociotechnical systems, and especially AI systems, are similarly dynamic and continual. Design is a process that continues post-deployment. This process of *redesign* for incongruencies, such as those emerging between system operations and various conditions for its compliance, need not *only* be undertaken by the designers themselves. Redesign can also involve users. This is of particular importance given that, in many cases, the designers of a system become detached from operations once it is deployed (i.e., one would be hard-pressed to imagine the designer(s) in a forward operating base). The AI-system's ability to learn, change, and adapt to dynamic scenarios makes this point even more salient, highlighting the need for continual monitoring over the operational lifecycle of a system.

5.3 Challenges posed by Artificial Intelligence (AI)

AI applications pose specific challenges when it comes to VSD, particularly in light of their self-learning capabilities mentioned already. These capabilities complicate the reliable integration of values in the design of AI technologies. To illustrate this, both van de Poel and myself employ a short, imaginary, and illustrative example of the complications raised by AI for VSD.

Suppose the tax department of a certain country wants to develop an algorithm to help detect potential cases of fraud. More specifically, the application should help civil servants select those citizens whose tax declaration needs extra or special scrutiny. Now, suppose they choose to build a self-learning artificial neural network for this task. An artificial neural network consists of a number of input units, hidden units, and one or more output units as pictured in Figure 3.

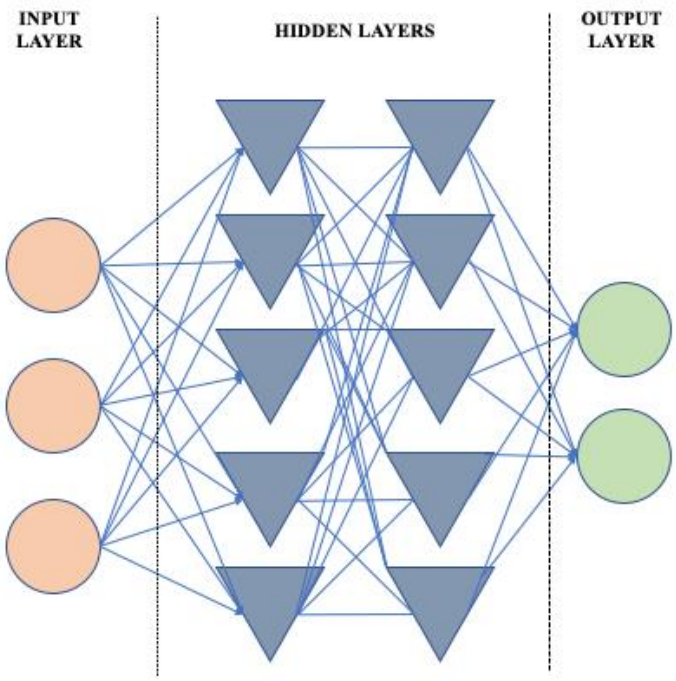

Figure 3. An artificial neural network (Source: Umbrello and van de Poel, 2021).

Suppose the output unit or variable is simply a yes/no indicating whether a specific tax declaration needs additional scrutiny. There can be many input variables (units), such as the amount of tax to be paid by a certain person, the use of specific tax exemptions, individual prior history (e.g., of suspected fraud in the past), and also personal details (age, sex, place of residence, etc.). Figure 3 shows how the units (variables) in the artificial neural network are connected. Connections between the units can be weighted factors learned by the algorithm. This learning can be supervised or not (Russell & Norvig, 2010). If supervised learning is applied, the algorithm may learn to make calls on which tax declarations need scrutiny – calls that are similar to those of experienced civil servants at the tax office. In the case of unsupervised learning, information on which scrutinised cases led to detection of actual fraud may be fed back into the algorithm. It can then be programmed to learn to select those cases that have the highest probability of leading to the detection of actual fraud (De Roux et al., 2018).

Now, one of the values that is obviously important in the design of such an algorithm is 'freedom from bias'. This value is already included on the original list of VSD values proposed by Friedman and Kahn, Jr., (2002). Friedman and Nissenbaum (1996, p. 332) define 'bias' in reference to "computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favour of others." In traditional VSD, this value may be implemented in the design of the

algorithm in a number of ways. First and foremost, it may be translated into design requirements that no variables in the artificial neural network (the nodes in Figure 3) use, as such variables may lead to an unwanted bias. Ethnicity, for instance, could be ruled out as a potential variable. But this will not be enough to ensure realisation of the value 'freedom from bias' as bias may also be introduced through proxy variables. Postal codes could be a proxy variable for ethnicity, so one may also want to rule out the use of such variables to ensure 'freedom from bias' (DeCamp & Lindvall, 2020; Kirkpatrick, 2016).

But even then, a self-learning algorithm could be biased due to the way it learns (Mehrabi et al., 2019). It may, for example, be biased because the training set for the algorithm is not representative or otherwise skewed. If a form of supervised learning is chosen, the algorithm could conceivably learn the biases already present in human judgments made for supervisory learning. Even if these potential sources of bias are also excluded, there is still no guarantee that the resulting algorithm is free of bias – certainly not if a form of non-supervised (reinforcement) learning is chosen. One issue is that that the resulting artificial neural network may be *described* as following a certain rule even if this rule was neither encoded nor (easily) derived from the nodes (variables) in the artificial neural network (c.f. Walmsley, 2012). In other words, the resulting algorithm can conceivably be described as following a rule that is somehow biased without this result being foreseeable or even clearly discernible to designers.

Bias in the algorithm of this imaginary case may thus be *emergent* and *opaque*. It is emergent in the sense of an unintended and unforeseen consequence from the way the algorithm has learned. It is opaque in the sense that the bias may not be immediately clear, to humans at least, from inspection of the algorithm or artificial neural network. This point is more general and does not just apply to this specific example or the value 'freedom from bias' (or 'fairness'). Due to their self-learning capabilities, AI systems – particularly those powered by ML – may develop features that were neither intended nor foreseen (or foreseeable) by their designers. They may have unintended value consequences. They may even unintentionally 'disembody' values embedded in their original design (van de Poel, 2020; Vanderelst & Winfield, 2018). Moreover, these unintended features may not always be discernible as they could derive from specific ways the algorithm has developed itself. These ways of learning may be hard, or even impossible, for humans to fully understand.

Such issues are not necessarily insurmountable. In the imaginary case of the algorithm for the tax office, technical solutions could make system development in a biased direction at least much more unlikely. We could tell the algorithm to optimise itself not only in terms of effectiveness (expressed in the number or percentage of cases of fraud detected, for example), but also in terms of fairness (such as

by presenting a non-biased selection of cases for investigation) (Mehrabi et al., 2019). The important point is that addressing emergence and opacity requires a set of design principles, or rather *design norms*, that are not needed for traditional technologies. Some of these principles relate to technical or design requirements, others to the organisation of the design process and the further life-cycle of a product (e.g., continued monitoring), and still others may have to do with which AI techniques to use or not. In Chapter 6, I propose the AI4SG principles as a way to address the specific challenges that AI poses to VSD.

5.4 Systems Engineering as *the* VSD Ontology

In Part I of this dissertation, I identified systems thinking (the verbial of the theoretical approach known as systems theory) as an apt framework for understanding both MHC of AWS as well as how to actually design AWS *for* MHC. There are multiple reasons for adopting this framework and they bear reiterating. The first reason for using systems theory is that it implicitly characterises the two levels of abstraction for understanding MHC discussed in the Chapters 3 and 4. The operational level of control is characterised by a plurality of actors and networks that complicates, yet also constitutes, how military operations are structured, planned, and carried out. Likewise, the design level of control is fundamentally built on the notion of tracking and tracing networks of systems and actors both in the use and in the design histories of those systems.

Secondly, systems theory is the theoretical framework from which systems engineering derives. As discussed in Chapter 2.3, systems engineering developed in the domain of defence. It is essentially the practical and managerial implementation of a systems thinking ontology. Aside from the obvious congruency between systems engineering and systems thinking within the military sphere, VSD maps onto systems-thinking design methodology (its underlying philosophical precepts, such as an *interactional stance* on technology, make this abundantly clear). VSD acknowledges that technology and societal forces co-construct and co-vary with one another, fundamentally affirming the *socialtechnicity* of systems (i.e., sociotechnical systems such as AI). This means that various actors, institutions, technologies, and their design histories form complex yet important networks of interaction. These relationships need to be brought to the fore in order for salient and responsible innovation to take place. Doing this means not overcomplicating this thesis with banal or unimportant theoretical constructs. Rather, the thesis intends to make manifest what is always already there: the fact that AI are not discrete technologies, but rather sociotechnical systems. By framing them as such, we can approach design *for* MHC of AWS more holistically.

5.4.1 The Sociotechnicity of AI Systems

In his 2020 paper *Embedding Values in Artificial Intelligence (AI) Systems*, Ibo van de Poel synthesises much of the literature on what constitutes sociotechnical systems and how such a framing can be used for understanding the particularities that distinguish AI systems from other sociotechnical systems. Here, I adopt much of this understanding for framing the particular instantiations of AI in the form of (fully) AWS. Like van de Poel, I understand sociotechnical systems as "systems that depend on not only technical hardware but also human behaviour and social institutions for their proper functioning" (c.f., Kroes et al., 2006). In this definition, sociotechnical systems are made up of three interrelated elements: (1) technical artifacts, (2) human agents, and (3) institutions (i.e., the norms followed by 2) (van de Poel, 2020, p. 391).

AI systems differ in that they not only possess all three of the above features, but also artificial versions of (2) and (3) within their architecture (1). These artificial varieties of (2) and (3) are called "artificial agents" and "artificial norms," respectively (van de Poel, 2020, p. 391). Intentionality is what distinguishes the human variety of (2) and (3) from their artificial parallels, which are characterised in physical-causal terms instead (see Table 1).[20]

	Intentional	**Physical-Causal**
Artifacts	Technical Artifacts	
Agents	Human Agents	Artificial agents
Norms	Institutions	Technical norms

Table 1. The basic building blocks of an AI system [*modified*] (Source: van de Poel, 2020, p. 391).

The left-hand column begins with technical artifacts, which are intentional in the sense that they are technical objects designed (intended) *for* operations toward certain functions (Kroes, 2010). The source of intentionality is not the artifact itself, of course, but the human agents. This does not mean that technical artifacts are devoid of intention. Rather, they are the bearers of intention obtained from

[20] Here, one should not be tempted to compare artificial agents with artificial moral agents (AMAs). AMAs are entirely eschewed here based on strong arguments against both the possibility of their development as well as their desirability/utility (Van Wynsberghe et al., 2019).

human agents (i.e., designers). Van de Poel further adopts the 'use plan characterisation' of technical artefacts described by Houkes and Vermaas (2010, c.f., Chapter 2) which characterises technical artifacts as aggregates of both physical structures and use plans (van de Poel, 2020, p. 391). The latter of these is described as the plan or guide for the projected proper use of an artefact as relates to the goal-realisation of its designed function(s) (Houkes & Vermaas, 2010, p. 28).

In terms of agents in Table 1, there are both intentional agents (i.e., human agents) as well as physical-causal agents, (i.e., artificial agents or AAs). AAs are what distinguish AI systems from other types of sociotechnical systems where humans are the agents in the above illustrated tripartite structure for sociotechnical systems (van de Poel, 2020, p. 391-2). The fundamental distinction between sociotechnical systems with AAs and those with none is that the former mirror many of the abilities possessed by human agents while not actually being human. Once again, AI technologies are defined as autonomous, interactive, adaptive, and capable of carrying out human-like tasks (Floridi & Sanders, 2004). However (and van de Poel is uncontroversially clear here), AAs should not be confused with moral agents in a sense similar to how we would describe human agents. There is no exclusive or exhaustive list of skills or features that can distinguish the agency of human agents from AAs. But this is all the more reason to not tokenise the agency of AI as part of the same type in the 'intentional' column. Rather, we should type it as something different, i.e., physical-causal (van de Poel, 2020, p. 392).

The third part that constitutes a sociotechnical system is *norms*. As with agents, norms are also differentiated into two types: intentional (i.e., institutions/social norms) and physical-causal (i.e., technical norms) (Bicchieri, 2005). In terms of human agents, these socioculturally situated, contingent institutions or social norms both support and constrain actions and decisions within certain contexts. Moreover, these types of norms can be explicit as well as tacit (Calvert, 1995). Because these types of norms are essentially social constructs, AAs cannot directly pattern themselves onto them as would be reasonably expected by human agents (van de Poel, 2020, p. 392). But the physical-causal architecture of AAs allows design for a technical counterpart of institutions or, in other words, *technical norms*.

5.4.2 Embodying Values in AI Systems

Van de Poel (2020) provides a thorough account for how each of the elements within Table 1 can and cannot embody values. It merits noting that for the systems level of AI (as opposed to simply the element level), van de Poel (2020) offers the following account as a more salient way to begin thinking about designing AI systems to embody values: "Value V is embodied in sociotechnical system S if S is

conducive to V because of those components of S that have been designed for V" (p. 403). Here, the conditional is neither overly exclusive nor exhaustive. It does not mandate that *all* components in an AI system need be designed for V in order for V to become embodied. This is because AI systems are often not (if at all) designed as whole systems. V can be obtained when relevant institutions and technical norms are likewise conducive to V. This notion of value embodiment permits, if not mandates, (re)design as a fundamental part of design practices. Relevant embodiable components of the system conducive to V for S suffice to embody V; a complete overhaul is not needed (van de Poel, 2020, p. 404). This account also allows for current sociotechnical systems already in pervasive use to embody values, which can change over time, through redesign. What components are relevant to any given system is predicated on the context. But given that institutions form the glue maintaining, supporting, and constraining the other components, van de Poel (2020) suggests they are a good starting point.

5.5 Conclusions

What this means for AWS is that the often-criticised unholy alliance between the military and industry (the military-industrial complex or MIC) becomes useful if such institutions are designed to be conducive to various values that critics of AWS argue are/will be lacking in fully AWS (i.e., conducive to the LOAC, IHL, etc.). But as described in Part I, the perhaps erroneous focus on autonomy as *mala in se* regarding AWS is likewise reiterated in this account of embodiment. The focus on autonomy has been a hallmark of research on AI. Yet this focus should fall less on AWS themselves and more on the norms, both institutional and technical, that mostly govern such autonomy. As I have aimed to argue here, VSD is fundamentally predicated on a systems-thinking approach. The embodiment of values is likewise supported and constrained by the many components that constitute such complex sociotechnical systems. However, the added elements that distinguish AI systems from other sociotechnical systems (i.e., AAs and technical norms) create additional loci to nest values within. The added complexity is mostly a function of the ability of AI to learn, adapt, and evolve over time, which further risks the disembodiment of values. Redesign as a function of full life-cycle monitoring (discussed in Chapter 6) provides a path to address these concerns and maintain MHC.

The following chapter proposes an adapted VSD approach as a means to address the challenges discussed in this chapter. In doing so, it proposes AI4SG factors as a starting point to bridge more abstract values with technical design requirements.

References

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., & Robinson, D. G. (2020). Roles for computing in social change. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 252–260.

Baum, S. D. (2016). On the promotion of safe and socially beneficial artificial intelligence. *AI and Society*, *July*, 1–9. https://doi.org/10.1007/s00146-016-0677-0

Bicchieri, C. (2005). *The Grammar of Society*. Cambridge University Press. https://doi.org/10.1017/CBO9780511616037

Borning, A., & Muller, M. (2012). Next steps for value sensitive design. *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*, 1125. https://doi.org/10.1145/2207676.2208560

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

Calvert, R. L. (1995). The rational choice theory of social institutions: cooperation, coordination, and communication. In *Modern Political Economy* (pp. 216–268). Cambridge University Press. https://doi.org/10.1017/CBO9780511625725.011

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

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

de Roux, D., Perez, B., Moreno, A., Villamil, M. del P., & Figueroa, C. (2018). Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 215–222.

DeCamp, M., & Lindvall, C. (2020). Latent bias and the implementation of artificial intelligence in medicine. *Journal of the American Medical Informatics Association*.

Flanagan, M., C. Howe, D., & Nissenbaum, H. (2008). Embodying Values in Technology: Theory and Practice. In J. van den Hoven & J. Weckert (Eds.), *Information Technology and Moral Philosophy* (pp. 322–353). Cambridge University Press. http://www.cambridge.org/catalogue/catalogue.asp?isbn=9780521855495&ss=cop

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-019-0055-y

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*, *26*(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349–379.

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 Jr, P. H. (2002). Human values, ethics, and design. In *The human-computer interaction handbook* (pp. 1209–1233). CRC Press.

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

Friedman, B., Kahn Jr., P. H., & Borinng, A. (2008). Value Sensitive Design and Information Systems. In K. E. Himma & H. T. Tavani (Eds.), *The Handbook of Information and Computer Ethics*. John Wiley & Sons, Inc. 10.1002/9780470281819.ch4

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330–347.

High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI | Shaping Europe's digital future*. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Houkes, W., & Vermaas, P. E. (2010). *Technical Functions: On the Use and Design of Artefacts*. Springer Netherlands. https://doi.org/10.1007/978-90-481-3900-2

Initiative, I. G. (2016). Ethically Aligned Design. *IEEE Standards V1*.

Jacobson, D. (2013). Wrong Kind of Reasons Problem. In H. Lafollette (Ed.), *International Encyclopedia of Ethics*. Blackwell Publishing Ltd. https://doi.org/10.1002/9781444367072.wbiee136

Kirkpatrick, K. (2016). Battling Algorithmic Bias: How Do We Ensure Algorithms Treat Us Fairly? *Commun. ACM*, *59*(10), 16–17. https://doi.org/10.1145/2983270

Kroes, P. (2010). Engineering and the dual nature of technical artefacts. *Cambridge Journal of Economics*, *34*(1), 51–62. https://doi.org/10.1093/cje/bep019

Kroes, P., Franssen, M., Poel, I. van de, & Ottens, M. (2006). Treating socio-technical systems as engineering systems: some conceptual problems. *Systems Research and Behavioural Science*, *23*(6), 803–814. https://doi.org/10.1002/sres.703

le Dantec, C. A., Poole, E. S., & Wyche, S. P. (2009). Values As Lived Experience: Evolving Value Sensitive Design in Support of Value Discovery. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1141–1150. https://doi.org/10.1145/1518701.1518875

Liao, Q. V., & Muller, M. (2019). Enabling Value Sensitive AI Systems through Participatory Design Fictions. *ArXiv Preprint ArXiv:1912.07381*.

Longo, F., Padovano, A., & Umbrello, S. (2020). Value-oriented and ethical technology engineering in industry 5.0: A human-centric perspective for the design of the factory of the future. *Applied Sciences (Switzerland)*, *10*(12), 1–25. https://doi.org/10.3390/APP10124182

Mabaso, B. A. (2020). Artificial Moral Agents Within an Ethos of AI4SG. *Philosophy & Technology*, 1–15.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *ArXiv Preprint ArXiv:1908.09635*.

Mok, L., & Hyysalo, S. (2018). Designing for energy transition through Value Sensitive Design. *Design Studies*, *54*, 162–183.

Mouter, N., de Geest, A., & Doorn, N. (2018). A values-based approach to energy controversies: Value-sensitive design applied to the Groningen gas controversy in the Netherlands. *Energy Policy*, *122*, 639–648.

Russell, S. j., & Norvig, P. (2010). *Artificial intelligence: A Modern Approach* (3rd Edition). Pearson. https://www.pearson.com/us/higher-education/program/Russell-Artificial-Intelligence-A-Modern-Approach-3rd-Edition/PGM156683.html

Sharkey, N. (2014). Towards a principle for the human supervisory control of robot weapons. *Politica & Societa*, *3*(2), 305–324.

Smith, R. C., & Iversen, O. S. (2018). Participatory design for sustainable social change. *Design Studies*, *59*, 9–36.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*(6404), 751–752.

Taylor, I. (2020). *Who Is Responsible for Killer Robots ? Autonomous Weapons, Group Agency, and the Military-Industrial Complex*. 1–15. https://doi.org/10.1111/japp.12469

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

Timmermans, J., Zhao, Y., & van den Hoven, J. (2011). Ethics and Nanopharmacy: Value Sensitive Design of New Drugs. *NanoEthics*, *5*(3), 269–283. https://doi.org/10.1007/s11569-011-0135-x

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

Umbrello, S. (2019). Atomically Precise Manufacturing and Responsible Innovation: A Value Sensitive Design Approach to Explorative Nanophilosophy. *International Journal of Technoethics*, *10*(2), 1–21. https://doi.org/10.4018/IJT.2019070101

Umbrello, S. (2020a). Combinatory and Complementary Practices of Values and Virtues in Design: A Reply to Reijers and Gordijn. *Filosofia*. https://www.academia.edu/43531194/Combinatory_and_Complementary_Practices_of_Values_and_Virtues_in_Design_A_Reply_to_Reijers_and_Gordijn

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

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

Umbrello, S., & van de Poel, I. (2021). Mapping Value Sensitive Design onto AI for Social Good Principles. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00038-3

UNESCO. (2017). *Education for Sustainable Development Goals: learning objectives*. United Nations Educational, Scientific and Cultural Organization,.

van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. *Minds and Machines*, *30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

van den Hoven, J. (2013). Architecture and Value-Sensitive Design. In C. Basta & S. Moroni (Eds.), *Ethics, design and planning of the built environment* (p. 224). Springer Science & Business Media. https://books.google.ca/books?id=VVM_AAAAQBAJ&dq=moral+value+such+as+freedom,+equality,+trust,+autonomy+or+privacy+justice+%5Bthat%5D+is+facilitated+or+constrained+by+technology&source=gbs_navlinks_s

van den Hoven, J. (2017). The Design Turn in Applied Ethics. In J. van den Hoven, S. Miller, & T. Pogge (Eds.), *Designing in Ethics* (pp. 11–31). Cambridge University Press. https://doi.org/10.1017/9780511844317

van den Hoven, J., Miller, S., & Pogge, T. (2012). The Design Turn in Applied Ethics. In Kenneth E. Himma & H. T. Tavani (Eds.), *The Handbook of Information and Computer Ethic*. Cambridge University Press.

van den Hoven, J., Vermaas, P. E., & van de Poel, I. (2015). Handbook of ethics, values, and technological design: Sources, theory, values and application domains. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Springer Reference*. Springer Netherlands. https://doi.org/10.1007/978-94-007-6970-0

van Wynsberghe, A. (2013). Designing Robots for Care: Care Centered Value-Sensitive Design. *Science and Engineering Ethics*, *19*(2), 407–433. https://doi.org/10.1007/s11948-011-9343-6

van Wynsberghe, A., Robbins, S., Robbins scott, S., van Wynsberghe, A., & Robbins, S. (2019). Critiquing the Reasons for Making Artificial Moral Agents. *Science and Engineering Ethics*, *25*, 719–735. https://doi.org/10.1007/s11948-018-0030-8

Vanderelst, D., & Winfield, A. (2018). The dark side of ethical robots. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 317–322.

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Colum. Bus. L. Rev.*, 494.

Walmsley, J. (2012). *Mind and Machine*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137283429

Whitman, M., Hsiang, C., & Roark, K. (2018). Potential for Participatory Big Data Ethics and Algorithm Design: A Scoping Mapping Review. *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2*. https://doi.org/10.1145/3210604.3210644

Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). Machine ethics: The design and governance of ethical ai and autonomous systems. *Proceedings of the IEEE*, *107*(3), 509–517. https://doi.org/10.1109/JPROC.2019.2900622

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

Woelfer, J. P., Iverson, A., Hendry, D. G., Friedman, B., & Gill, B. T. (2011). Improving the Safety of Homeless Young People with Mobile Phones: Values, Form and Function. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1707–1716. https://doi.org/10.1145/1978942.1979191

6 Adapting the VSD Approach

When people know a number of things, and one of them understands how the things are systematically categorised and related, that person has an advantage over the others who don't have the same understanding — Luzatto (circa 1735)

6.1 AI for Social Good: Norms for AI Design

The most thorough work on the harmonisation of AI4SG values was recently undertaken by Cowls, King, Taddeo, & Floridi (2019), who focus on factors 'particularly relevant' to AI (i.e., not exhausting the potential list of relevant factors). The seven factors that are particularly relevant for the design of AI towards social good are: (1) *falsifiability and incremental deployment*; (2) *safeguards against the manipulation of predictors*; (3*) receiver-contextualised intervention*; (4) *receiver-contextualised explanation and transparent purposes*; (5) *privacy protection and data subject consent*; (6) *situational fairness*; and (7) *human-friendly semanticisation* (Floridi et al., 2020, p. 1773).

Although discussed separately, the seven factors naturally co-depend and co-vary with one another. Thus, they should not be understood as a rank-ordered hierarchy. These factors further relate, in some way, to at least one of the four ethical principles that the EU High-Level Expert Group on AI lays out: *respect for human autonomy*, *prevention of harm*, *fairness* and *explicability*. This mapping onto the more general values of ethical AI is not insignificant, as any divergence from these more general values has potentially deleterious consequences. What the seven factors are meant to do, then, is specify these higher-order values as more specific norms and design requirements (Figure 1).

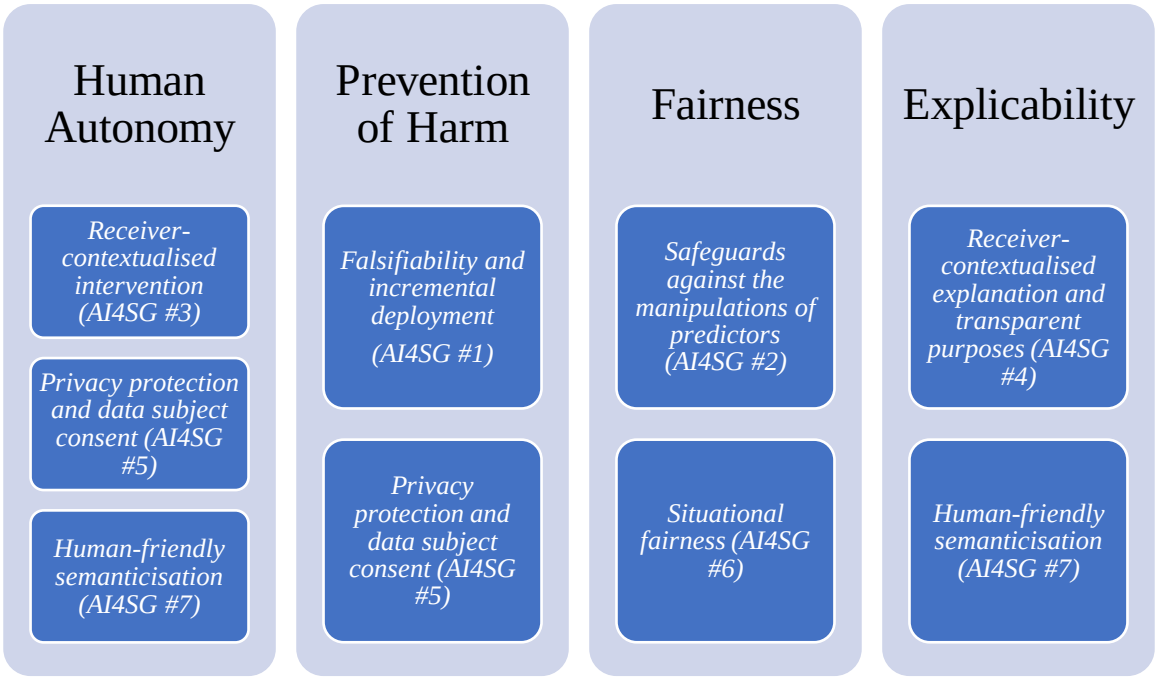

Figure 1. Relationship between higher-order values of the EU HLEG on AI and AI4SG norms.

Rather than reiterate what has already been clearly evaluated and discussed by Floridi et al. (2020), the below paragraphs briefly summarise each of the seven factors (later discussed in more detail) alongside ways that VSD practices can be levied to actualise these factors.

(1) *Falsifiability and incremental deployment.* To move the development of AI forward towards the embodiment of values such as transparency and safety, the value of *falsifiability* is important. This is because it is considered a critical factor in the social acceptance and trust of technologies more broadly. Falsifiability is defined as "the specification, and the possibility of empirical testing, of one or more critical requirements, that is, an essential condition, resource, or means for a capability to be fully operational, such that something could or should not work without it" (Floridi et al., 2020, p. 1777). Other values implicated in AI design are thus predicated on their ability to be falsifiable or essential to the architectures of a technical system.

This entails continued empirical testing, which must be undertaken in different contexts (and obviously cannot be exhausted without full deployment of a system) to best ascertain the possible failures for a system. There is thus a need for an incremental deployment cycle wherein systems are introduced into real-world contexts only when a minimum level of safety makes such deployment warranted. In sum, "AI4SG designers should identify falsifiable requirements and test them in incremental steps from the lab to the 'outside world'" (Floridi et al., 2020, p. 7).

(2) *Safeguards against the manipulations of predictors*. The manipulation of predictors can lead to a range of potentially deleterious outcomes for AI, moving away from the promises of AI4SG. Floridi et al. (2020) describe the outcome of the manipulation of input data as well as overreliance on non-causal indicators (p. 1779). The nature of overreliance on non-causal indicators as well as the often overespoused but underthought value of transparency can lead to the gamification of systems towards desired ends by those who understand what inputs lead to what outputs (c.f., Boscoe, 2019; Ghani, 2016). To avoid this, Floridi et al. (2020) argue that "AI4SG designers should adopt safeguards…[to] ensure that non-causal indicators do not inappropriately skew interventions, …[and] limit, when appropriate, knowledge of how inputs affect outputs from AI4SG systems, to prevent manipulation" (Floridi et al., 2020, p. 1779).

(3) *Receiver-contextualised intervention.* The co-construction and co-variance of technologies and users implicates a delicate balancing act between artifacts and their effect on user autonomy. Within the context of technological design and development, this is a value of particular importance (Umbrello, 2019b). To balance the false positives and negatives that can result in suboptimal levels of user-technology interventions, users can be given *optionality*. This provides one possible route for balancing interventions on autonomy. Optionality is contextualised based on "information about users' capacities, preferences and goals, and the circumstances in which the intervention will take effect" (Floridi et al., 2020, p. 1780). Briefly,

> AI4SG designers should build-decision-making systems in consultation with users interacting with and impacted, by these systems; with understanding of users' characteristics, of the methods of coordination, and the purposes and effects of an intervention; and with respect for users' right to ignore or modify interventions. (Floridi et al., 2020, p. 1780)

(4) *Receiver-contextualised explanation and transparent purposes.* The aims of any given system must be transparent. In other words, operations carried out by a system should be explicable or explainable so as to be understood. Given that the intricacies of the operations and objectives of a system are the consequence of design decisions, design is inextricably linked to these values. The evermore ubiquitous deployment of AI systems is already underway. The need for explicability and transparency in their operations and goals has garnered a lot of attention due to the potential harm that can come about as a consequence of opaque goals and operations (Allo et al., 2016; Turilli & Floridi, 2009). In terms of (3), the information used to explain the operations and objectives of a system should also be receiver-contextualised (Floridi et al., 2020).

Because the goals, design programs, and tools used for differing AI4SG projects vary greatly, correct contextualisation will similarly vary. Floridi (2017) calls this conceptual schema (of what is

being framed for whom) the Level of Abstraction. The Level of Abstraction consists of the five components that comprise any theory of a given system[21]. Because the inner workings and overall goals of any AI system are the outcomes of designer choices and design flows, there must be transparency regarding design decisions to determine if they map onto the motivation behind the design and deployment of any given system. The type of transparency, the goals, and designer intentions along with the level of transparency needed for successful explicability of the operations and goals of AI systems must necessarily be determined in the early stages of the design program in question. In other words,

> AI4SG designers should choose a Level of Abstraction for AI explanation that fulfils the desired explanatory purpose and is appropriate to the system and the receivers; then deploy arguments that are rationally and suitably persuasive for the receivers to deliver the explanation and ensure that the goal (the system's purpose) for which an AI4SG system is developed and deployed is knowable to receivers of its outputs by default. (Floridi et al., 2020, p. 1784)

(5) *Privacy protection and data subject consent.* Scholarship on privacy protection and subject consent is both rich and nuanced, encompassing decades of socio-ethical and legal perspectives (among others) informing these topics. Given that privacy forms the basis for both good policy and just democratic regimes (Peters, 2018), AI4SG programs should naturally make this an essential factor (Solove, 2008). Tensions and boundaries between different levels and understandings of user data processing and use have already been explored; moreover, nuances in terms of how to adequately address such tensions have been proposed (Floridi, 2016; Price & Cohen, 2019). As stakeholder data is foundational to the usability and efficacy of AI systems, AI4SG systems must seek to provide a sufficient balance that respects the values of stakeholders in regard to data processing and storage. Accordingly, Floridi et al. (2020) note that "AI4SG designers should respect the threshold of consent established for the processing of datasets of personal data" (p. 1786).

(6) *Situational fairness*. As mentioned in (5), data sets are critical to the function of AI systems. Datasets themselves can be biased on account of multiple factors (dataset collection, selection, categorisations, etc.). The resulting function of any given system can thus provide biased results (Boscoe, 2019). Biased decision-making can take on ethical importance when relevant datasets involve ethically relevant categories for data, such as race, gender, or age, among other possibilities (Friedman & Nissenbaum, 1996). If we are to attain AI4SG, the propagation of bias in datasets must be avoided.

[21] For the sake of brevity and conciseness, I do not include the full description of the five levels of Abstraction. For further on this, we direct the reader to Floridi, L. 2017. The logic of design as a conceptual logic of information. Minds Mach. 27, 495–519.

This is because recursive improvements to systems only exacerbate bias if such improvements are designed or trained using biased datasets. So, "AI4SG designers should remove from relevant datasets variables and proxies that are irrelevant to an outcome, except when their inclusion supports inclusivity, safety, or other ethical imperatives" (Floridi et al., 2020, p. 1788).

(7) *Human-friendly semanticisation*. Managing and maximising the 'semantic capital' of agents must be essential to the design of AI4SG systems. Floridi (2018) defines semantic capital as "any content that can enhance someone's power to give meaning to and make sense of (semanticise) something" (p. 483). AI allows for automation of semanticisation, i.e., making sense of things, which can lead to ethically problematic results if done haphasardly. Arbitrary semanticisation can give meaning in ways that do not map onto our own understandings (random meaning-making). AI semanticisation can also be too narrow due to limited dataset exposure that allows for propagation of similarly narrow meanings, thus limiting the redefinition or interpretation of things (Al-Abdulkarim et al., 2016). Semanticisation is subjective due to the fact that the agent engaging in semanticisation is essential to what and how meaning is made. AI systems aimed at total semanticisation are thus unworkable and quixotic. The way around this is to delimit the tasks carried out by AI systems. There need not be a total abdication of tasks. Rather, the ones that must be carried out by AI systems should be determined *a priori* to the deployment of an AI4SG system (Floridi et al., 2020). For this reason, "AI4SG designers should not hinder the ability for people to semanticise (that is, to give meaning to, and make sense of) something" (Floridi et al., 2020, p. 1789).

This section has condensed the seven essential factors necessary to the design of AI4SG systems as proposed by Floridi et al. (2020). We can now see how these factors help overcome the challenges posed to VSD by AI discussed in the previous chapter. If we adopt the specific example given in Chapter 5, then AI4SG norm #6 would require "[removing] from relevant datasets variables and proxies that are irrelevant to an outcome" (Floridi et al., 2020, p. 1788). This is in line with the traditional VSD approach, but it is not enough as AI bias may be emergent and/or hidden (opaque). To address the emergent character of bias, norm #1 is particularly important due to the emphasis on incremental development. This is primarily a procedural requirement that requires monitoring and extending VSD to the full-life cycle of design, as we discuss in greater detail in section 6.4. To avoid opaqueness, AI4SG principles #4 and #7 are important. Sometimes, they may imply that certain ML techniques should not be used.

Taking these factors into consideration, the following sections integrate the discussion from the preceding chapter on VSD in an attempt to provide some preliminary approaches to employing VSD towards AI4SG design.

6.2 Integrating AI4SG Principles as Design Norms

To address the challenges posed to VSD by AI, we propose an adapted VSD approach. The adaptations we propose are threefold: 1) integrating AI4SG principles into VSD as design norms from which more specific design requirements can be derived; 2) distinguishing between values promoted by design and values respected by design to ensure that the resulting design not only does no harm, but also contributes to doing good; and 3) extending the VSD process to encompass the whole lifecycle of an AI technology in order to be able to monitor unintended value consequences and redesign the technology if necessary. This section briefly explains these new features, then sketches the overall process.

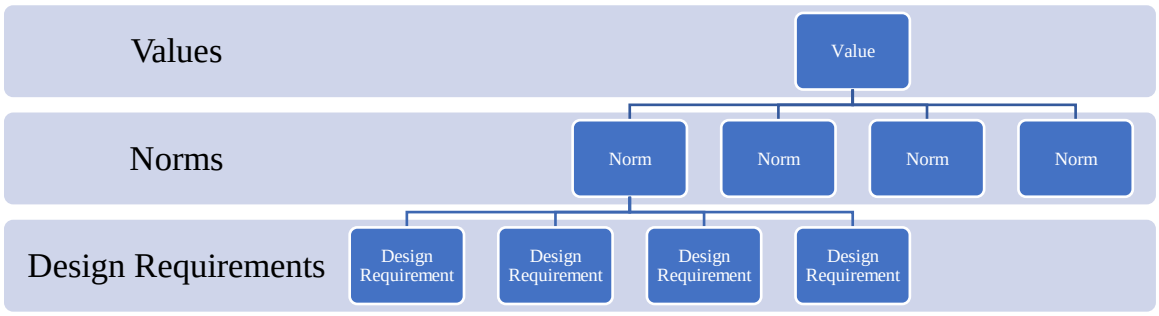

Figure 2. Values hierarchy. (Source: van de Poel, 2013)

6.3 Distinguishing between Values to be Promoted and Values to be Respected

In order for a VSD approach to AI to achieve more than just avoiding harm, an explicit orientation toward socially desirable ends is necessary. Such an orientation is still missing from current proposals for AI4SG. I propose addressing this gap with an explicit orientation to the Law of Armed Conflict (LOAC). This body of law, the LOAC, is predicated on the customary rules of International Humanitarian Law (IHL). It is the best approximation of what we collectively believe to be valuable military/wartime institutions.

When it comes to the control of AWS (*Jus ad Bellum*) as well as their deployment (*Jus in Bello*), the former are related to *a priori* international norms of conflict management and customary law

such as those of the UN Charter. The latter, which focuses on the rules of hostility *in situ*, include regulation such as the Geneva (in this case, Article 36 of Protocol I in particular [see below]) and Hague Conventions regarding the means and methods for hostilities. Figure 3 lists the LOAC and Article 36 as higher-order values to be promoted as much as possible.

In Chapter 3's discussion on the operational level of control, we saw how rigorous the principles and practices of military operations are prior to deployment and engagement. The LOAC are equally rigorous and considered inseparable from the customary law of each nation-state signatory. They are thus valid at all times. For this reason, each participant within a systems' operations is aware of these fundamental principles. As such, they are expected to become a frame for planning, practices, training, and operations.

Figure 3. Article 36[22] as the value to be respected regarding the design, engineering, and deployment of an AWS as a function of the six LOAC.

[22] In the study, the development, acquisition, or adoption of a new weapon, means, or method of warfare involves a High Contracting Party. The Party has an obligation to determine whether its employment would, whether in some or all

Now, for the sake of transparency, there are strong philosophical and legal arguments against the development of (fully) lethal autonomous weapons (LAWS)[23]. Gabriel Wood (2020) makes cogent arguments that the pro-LAWS position is self-undermining for various reasons. But he is also clear that such a position does not, and perhaps should not, exclude the possibility of researching and designing certain types of AWS. That is the position forwarded by this thesis at the onset (i.e., a more nuanced distinction between types of AWS to strengthen prohibitive measures). Proponents suggest LAWS are technically more capable, either now or in the future, of engaging in warfare to a higher degree of accuracy, speed, and target discrimination. The primary argument here is that if this is true, then such can also be said for their ability to conduct warfare in a non-lethal capacity. This nullifies the arguments for lethality as a necessary component. I, myself, have made the argument elsewise that such technical capacities make LAWS preferable (Umbrello, 2019a; Umbrello et al., 2020). For me, Wood's arguments are philosophically robust enough to neutralise them.

Still (and as this thesis proposes), the argument is less so about the problem of *lethality* and more so about the 'problem' of *autonomy* as the basis for a ban. His argument is oriented towards those of AWS proponents whereas mine is oriented towards those of anti-AWS or pro-ban proponents. But these seemingly opposite positions arrive at a similar conclusion. At the very least, Wood (2020) and I agree that:

> we should welcome the development of autonomous weapons while doing our utmost that they are programmed in such a way as to adhere to *all* the laws of war, bearing in mind how their own capabilities will affect the moral and legal prescriptions in a given scenario. (p. 234)

This is an attempt to do just that.

6.4 Extending VSD to the Entire Lifecycle

To address the emergent and possibly unintended properties acquired by AI systems as they learn, VSD should be extended to the full life cycle of AI technologies. This allows continued monitoring of the potential for unintended value consequences, which would further require redesigning the technology as needed (De Reuver et al., 2020; van de Poel, 2020). As already mentioned, AI4SG Principal (1) voices a similar idea: "AI4SG designers should identify falsifiable requirements and test them in

circumstances, be prohibited by the Protocol or any other rule of international law applicable to the High Contracting Party (Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (AP I), 1979).

[23] Recall the distinction (noted in the introduction and Annex I) that AWS can be categorised into offensive and defensive types, each of which can be both lethal and/or non-lethal depending on the munitions employed.

incremental steps from the lab to the 'outside world'" (Floridi et al., 2020, p. 1777). The need for ongoing monitoring arises from the uncertainties accompanying new technologies upon their introduction to society (van de Poel, 2016). The previous chapter discusses how this is a fundamental necessity in the precept of embodied values, as such values can manifest themselves in different ways throughout the lifecycle of an AI system (c.f., Mökander and Floridi (2021)). In these cases, *post hoc* redesign likewise becomes necessary. But it can only be triggered through epistemic access granted for continual monitoring over the full life cycle of the system.

6.5 Mapping Value Sensitive Design onto AI for Social Good Principles

Taking the above into account, VSD for AI proceeds in four iterative phases (Figure 4) briefly described below:

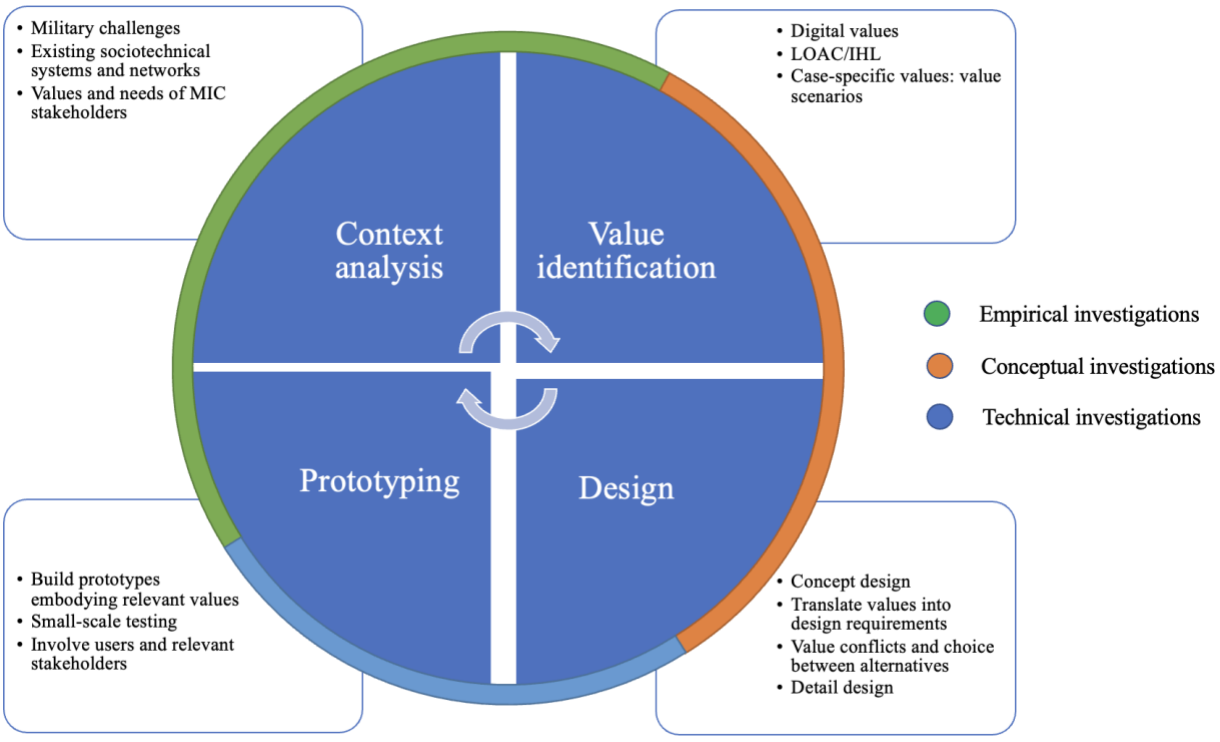

Figure 4. VSD design process for AI technologies. Source: Umbrello & van de Poel, 2021 (modified).

6.5.1 Context Analysis

Motivations for design differ across different projects. For this reason, there is no normative starting point from which all designers should begin. VSD acknowledges that technology design can begin with the discrete technology itself as a starting point, or the context for use, or a certain value (Figure 5). In all cases, analysis of the context is crucial. Various contextual variables come into play that impact the way values are understood (in the second phase), both in conceptual terms as well as in practice, on account of different sociocultural and political norms. Eliciting stakeholders in sociocultural contexts is imperative within the VSD approach to determine whether the explicated values of the project map faithfully onto those of both direct and indirect stakeholders. Empirical investigations thus play a key role in determining the potential boons and downfalls to any given context.

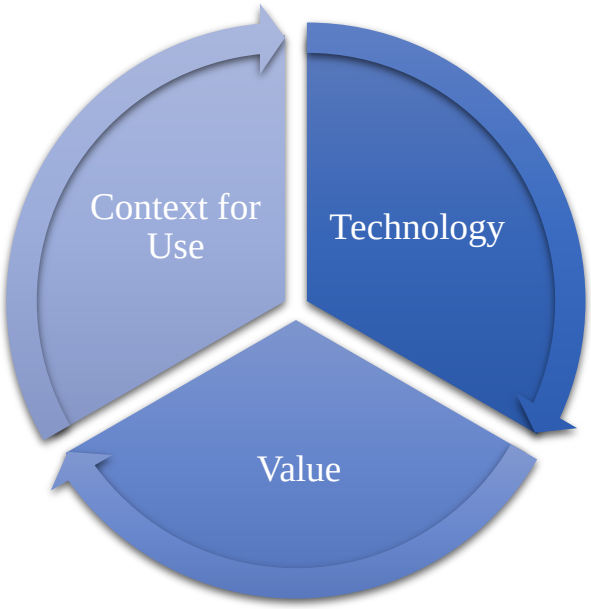

Figure 5. Starting considerations for VSD. Typically, one of the three is most pertinent to any given design. (Umbrello, 2021)

6.5.2 Value Identification

The second phase concerns identification of a set of values that form the starting point of the design process. We suggest three main sources for such values:

1) values that are to be promoted by the design, such as by deriving from the LOAC;
2) values that should be respected, especially those identified in relation to AI. These include *respect for human autonomy*, *prevention of harm* (nonmaleficence), *fairness,* and *explicability* (Floridi et al., 2018; High-Level Expert Group on AI, 2019); and

3) context-specific values that are not covered by the first two sources. They derive instead from analysis of the specific context in the first phase, especially of the values held by stakeholders.

It should be noted that the second phase does not just involve empirical investigations. Rather, it has a distinct normative flavour in the sense that it results in the identification of values that should be upheld in further design *from a normative point of view*. In addition, this phase involves conceptual investigations geared at interpreting (in context) and conceptualising relevant values.

6.5.3 Formulating Design Requirements

The third phase involves the formulation of design requirements on the basis of the contextual analysis (phase 1) and identified values (phase 2). Here, tools such as the value hierarchy can be useful to mutually relate values and design requirements or to translate values into design requirements (Figure 2). We suggest that the translation of values into design requirements is somewhat different for the sets of values formulated in the second phase. The first set of values derived from the LOAC and Article 36, for example, are values to be promoted. They are typically translated into design requirements formulated as criteria that should be achieved as much as possible. The second set of values are those that need to be respected, especially as relates to AI. We find the AI4SG principles are particularly helpful for formulating more specific design requirements. These requirements will most likely be formulated as constraints or boundary conditions rather than as criteria that should be achieved as much as possible; boundary conditions set the deontological constraints that any design must meet to be ethically (minimally) acceptable. For the third set of contextual values, the context analysis – and in particular the stakeholder analysis – will most likely play an important role in how these are translated into design requirements.

6.5.4 Prototyping

The fourth phase is the building of tests for prototypes that meet the design requirements. This idea is in line with what is more generally described in VSD as a "value-oriented mock-up, prototype, or field deployment" (Friedman & Hendry, 2019, p. 62). We propose extending this phase to the entire life cycle of an AI technology because, even if such technologies initially meet value-based design requirements, they may develop in such a way that unexpected and undesirable effects materialise. They could also simply no longer achieve the value for which they were intended, or their use may have unforeseen side effects that require consideration of additional values (van de Poel, 2018). In such cases, there is reason to redesign the technology and do another iteration of the cycle.

6.6 Conclusions

Predicated on the philosophical underpinnings of Chapter 5, this chapter outlines how VSD methodology can and should be adapted to meet the specific challenges that come with AI systems design. In order to ensure adoptability and illustrate the efficacy of this approach, the following chapter uses the example central to this thesis – aerial (fully) AWS – to more clearly show how the process works by situating it in a figurative context for a specific AI system.

References

Al-Abdulkarim, L., Atkinson, K., & Bench-Capon, T. (2016). A methodology for designing systems to reason with legal cases using abstract dialectical frameworks. *Artificial Intelligence and Law*, *24*(1), 1–49.

Allo, P., Taddeo, M., Floridi, L., Wachter, S., & Mittelstadt, B. D. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 205395171667967. https://doi.org/10.1177/2053951716679679

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

De Reuver, M., van Wynsberghe, A., Janssen, M., & Van de Poel, I. (2020). Digital platforms and responsible innovation: expanding value sensitive design to overcome ontological uncertainty. *Ethics and Information Technology*, 1–11.

Floridi, L. (2016). On human dignity as a foundation for the right to privacy. *Philosophy & Technology*, *29*(4), 307–312.

Floridi, L. (2017). The logic of design as a conceptual logic of information. *Minds and Machines*, *27*(3), 495–519.

Floridi, L. (2018). Semantic capital: Its nature, value, and curation. *Philosophy & Technology*, *31*(4), 481–497.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*, *26*(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

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

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330–347.

Gabriel Wood, N. (2020). The Problem with Killer Robots. *Journal of Military Ethics*, *19*(3), 220–240. https://doi.org/10.1080/15027570.2020.1849966

Ghani, R. (2016). *you say you want transparency and interpretability?* Blog Entry.

High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI | Shaping Europe's digital future*. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Mokander, J., & Floridi, L. (2021). Ethics-Based Auditing to Develop Trustworthy AI. *Minds and Machines*. https://doi.org/10.1007/s11023-021-09557-8

Peters, B. G. (2018). *Policy problems and policy design*. Edward Elgar Publishing.

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, *25*(1), 37–43.

Solove, D. J. (2008). *Understanding privacy* (Vol. 173). Harvard university press Cambridge, MA.

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*(2), 105–112. https://doi.org/10.1007/s10676-009-9187-9

Umbrello, S. (2019a). Lethal Autonomous Weapons: Designing War Machines with Values. *Delphi - Interdisciplinary Review of Emerging Technologies*, *2*(1), 30–34. https://doi.org/10.21552/delphi/2019/1/7

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

Umbrello, S. (2021). Conceptualizing Policy in Value Sensitive Design: A Machine Ethics Approach. In S. J. Thompson (Ed.), *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence* (pp. 108–125). IGI Global. https://doi.org/10.4018/978-1-7998-4894-3.ch007

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

Umbrello, S., & van de Poel, I. (2021). Mapping Value Sensitive Design onto AI for Social Good Principles. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00038-3

Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (AP I), Pub. L. No. 17512, 1125 (1979). https://treaties.un.org/doc/publication/unts/volume 1125/volume-1125-i-17512-english.pdf

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

van de Poel, I. (2016). An Ethical Framework for Evaluating Experimental Technology. *Science and Engineering Ethics*, *22*(3), 667–686. https://doi.org/10.1007/s11948-015-9724-3

van de Poel, I. (2018). Design for value change. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-018-9461-9

van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. *Minds and Machines*, *30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

7 The AI4SG-VSD Design Process in Action: Multi-Tiered Design and Multi-Tiered MHC

Autonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force.
— Department of Defence Directive 3000.09 (2012, p. 2)

7.1 Contextual Analysis

As discussed in the previous chapter and Annex II, a VSD program can begin in at least one of three ways: with (1) a technology, (2) one or more values, and/or (3) the context of use. In the case of AWSs, the context of use can be construed as the motivating factor behind their design and development, such as the need to extricate more human operators from hot zones and/or increase cost-efficiency while maintaining operational efficacy. The design of systems such as aerial fully AWSs should be explicitly oriented at trying to manage tensions and eliminate the moral overload of *prima facie* conflicting values (ICRC, 2002). The prioritisation and subsequent operationalisation of certain values over others is directly contingent on the context of use. Consequentially, this context can result in one or more values being set aside in favour of prioritising others. For example, the value of *discrimination* (in the sense of properly discriminating between targets, or *hors de combat*) may be set aside in favour of that of *military necessity* and *proportionality* with regards to aerial AWSs, since *discrimination* is a necessary part of the *a priori* briefings of the operational level of control and therefore may not need to be prioritised as a technical design requirement, whereas other values may (this may even include human judgment of the level of lethality with regards to the former two values as a function of weaponeering).

7.2 Value Identification

Figure 1. Three sources of values for the VSD of AI4SG (source: Umbrello et al., 2021)

7.2.1 Values to be promoted by design: LOACs

As mentioned in Chapter 5, AWSs, whether fully (level 5) or semi-autonomous, should not be construed as artificial moral agents (AMAs; ICRC, 2002) similar to human combatants, but rather as novel weapons that are capable of being designed so as to *embody* values, making them artificial agents (cf., Chapter 5). In light of this, Article 36 *de facto* brings these types of systems under its normative umbrella. Therefore, for such systems to be compliant and *embody* the values expressed by the LOACs, the design requirements translated from these higher-order values must promote compliance towards these LOACs as much as possible, despite their status as existent necessities that define much of the *operational level* of MHC (cf., Chapter 6, Figure 3).

> *Distinction* (GP I, Art. 48 and GP I, Art. 52 (2)): The MIC must clearly distinguish between civilian objects (i.e., people and associated public/private entities and infrastructures). To this end, civilians must be protected as much as possible, although they may of course lose this protection if they engage in hostilities.

> *Proportionality* (HR IV, Arts. 22 & 23 and GP I, Art. 57): When military targets are attacked, civilians must be protected as much as possible (i.e., *distinction*). Although civilian objects may

be collateral damage during engagement with a military target, such damage must be proportional and not excessive to the military objective; “To avoid violating this principle requires thought and effort. Poor planning and intelligence, slack staff work, leadership, command and control can easily result in the destruction of a whole town or village, with its hospitals, religious centres and civilian population” (ICRC, 2002, p. 12). Such proportionality assessments belong firmly in the *operational level* of MHC, as described in Chapter 3.

Military necessity (H IV, Art. 23 (g)): Contingent on the above two laws, military necessity allows for the realities of battle to manifest as permitting whatever “reasonable force is necessary, is lawful and can be operationally justified in combat to make your opponent submit” (ICRC, 2002, p. 12).

Limitation (HR IV, Arts. 22 & 23 and GP I, Arts. 35 & 57): In direct relation to AWSs, the means and tools used by states to wage warfare are not unconstrained. International Humanitarian Law (IHL) constrains if/how tactics and weapons are employed on the battlefield; “Weapons and tactics that are of a nature to cause unnecessary suffering or superfluous injury are prohibited” (ICRC, 2002, p. 12). As Gabriel Wood (2020) explains, AWSs that are sufficiently advanced, as suggested by their proponents, may *de facto* contravene the law of limitation, given their potential ability to be unimaginably fast, precise, efficient, and thereby *a fortiori* unlawful since the need for lethality consequentially becomes “unnecessary.”

Good Faith and *Humane Treatment and Non-Discrimination* as part of the LOACs are not discussed here because they are particular not to the technology of use but to the conduct of the military and their a priori use of such tools, rather than the technical function and design of the tools themselves. That is not to say that they are not important; however, for the purposes of this thesis, and to demonstrate how engineers can begin to think systemically about AWS designs that incorporate MHC, the four LOACs above are used as illustrations.

7.2.2 Values respected by the design

This second level of values are ones to be promoted, especially in relation to AI.

Respect for Human Autonomy: We increasingly interact with autonomous decision-making systems in different domains. Such systems influence our lives in various and multifaceted ways, from shaping the

context in which individual decision making occurs, to altering interactions between individuals and assumptions of democratic participation. Autonomy thus refers to the capability of agents to retain full freedom of choice, in tandem with the delegation of decisions to systems. Systems, in turn, should be designed so as to promote autonomy, avoiding those cases whereby their efficacy falls short in terms of making consistent and coherent decisions on the behalf of human users (Floridi et al., 2018). With regards to aerial fully AWSs, this autonomy is understood to mean both technical autonomy as a function of reason-responsiveness (i.e., the *design level* of MHC) and the constraints on autonomy as a function of the *operational level*.

Prevention of harm (or nonmaleficence): This value seeks to prevent risks and harm by the understanding the capabilities and limitations of the systems. This is of course in direct relation to both abstraction levels of MHC. More specifically, at least one (human) moral agent(s) must understand not only the systems' capabilities and limitations (i.e., on the design level), but also as the basis for weaponeering the AWS as a viable and therefore lawful option for any given operation. As mentioned in the previous chapter, this value should not be misconstrued as "doing no harm" in the most exclusive and exhaustive of senses. If such AWSs *actually* arrive at the technical capabilities that Arkin (2008), Guetlein (2005), and even myself (Umbrello et al., 2020) have espoused, then they would *de facto* violate the law of limitation, given that such technical prowess would make lethality *per se* unnecessary and subsequently unlawful. Although we have discussed machines designed for death, we must nonetheless remain lawful and always under MHC. This can only be accomplished if the agents involved in the design and deployment of such systems are sufficiently cooperative to ensure that the knowledge transfer between design and operation does not leave epistemic gaps. To this end, and similar to how van Wynsberghe (2012, p. 111) describes nonmaleficence in her care-centred framework for VSD, nonmaleficence can be subsumed under the value of *competence*, which asserts a system's capacities and limits regarding its task-engagement abilities. The capabilities of these systems may include safety, efficiency, and quality of task execution, among others.

Floridi et al. (2018) argue that the value of *fairness* can be framed as justice and defined in a tripartite manner: (1) using AI to correct past wrongs, such as eliminating unfair discrimination; (2) ensuring that the use of AI creates benefits that are shared (or at least sharable); (3) preventing the creation of new harms, such as the undermining of existing social structures (i.e., LOACs/the IHL). With regards to AWSs, fairness can be understood technically as promoting *distinction* to ensure lawful target acquisition and its subsequent weapons release, in addition to the social structure of the MIC to

promote *non-discrimination* more generally by not introducing technical systems that undermine this LOAC as it is more broadly understood.

The value of *explicability* means that AI systems should be intelligible and non-opaque, and there should be at least one agent that can be considered accountable for the operation of the system; i.e., the tracking condition of the *design level* (Floridi et al., 2018). In AI contexts, this raises questions regarding potential accountability gaps, which – beyond the issues of information disclosure and visibility – address the need for modalities to render systems explainable and understandable to users and stakeholders at large (Mecacci & de Sio, 2019; Pasquale, 2017).

7.2.3 Context-specific values not covered by (1) and (2)

As discussed in the introduction, Annex I, and Part I, the development of AWSs can be understood as being motivated by a number of converging factors, such as the existent trend of increasing automation and the efficiency provided by such systems, the military advantage of having tools that provide asymmetric gains, and, of course, the abdication of traditionally dangerous military operations away from human operators. Many of the values and ethical issues that have potentially emerged as a consequence of the design and deployment of AWSs have already been discussed, such as the values espoused by the LOACs and Article 36 governing new weapons technologies. However, if the two-tiered understanding of MHC described in this thesis is to *actually* obtain, then, as mentioned, the various industrial partners responsible for much of the design and engineering of AWSs cannot be extricated from the deployment process; rather, they need to form stronger partnerships as part of the MIC. This is the case since the design side of MHC (i.e., the industrial partners) forms a set of relevant and even direct stakeholders, seeing as the AWS should necessarily be reason-responsive, and also because the design histories of such systems necessarily trace back to these agents. They too must then be considered direct stakeholders for salient design to take place, rather than be allowed to abdicate their moral responsibility to the more direct operational and causal (albeit not always merely causal) forward-operating commanders on the deployment side. As such, given the role of the industrial side as fundamentally constitutive of MHC over these systems in action, in addition to the framework for MHC and how VSD can be employed *for* MHC (i.e., full-lifecycle monitoring and redesign), a contextual analysis serves to elicit such classes of values, which relate to the stakeholders' values and preferences (cf., the common interest of the international communities to preserve the values discussed in §7.1).

Aside from the higher-level distal values of the international community that have given rise to the LOACs, Article 36, and other norms that govern AWSs, the systems need to not only cater and thus be sensitive to this class of reasons (i.e., stakeholders) but also be responsive to distal and proximal reasons of the MICs, as they relate more closely to the military objectives and the fundamentally economic values that drive the industrial side of the network. In §7.3, I provide some examples of how MIC partnerships, particularly at the systems engineering level (i.e., the industrial side), can begin translating higher-order values such as the LOACs as well as the values important to the continued sustainability of their MIC partnerships as well as their own economic reasons. The latter values are not to be construed as being in tension with the other classes of reasons (i.e., economic values may only be *prima facie* in a state of moral overload (van den Hoven et al., 2012)) or with the LOACs/Article 36; instead they can be remediated through both salient engineering and regulatory measures that legitimise the group agency and moral responsibility of the MIC, thereby highlighting the existing relationship between these actors. As previously stated, the existent network that constitutes the MIC makes a distinction without differences, since the systemic synergy that results in the design and deployment of AWSs, like any other system-within-a-system in the military, is not isolatable, and any attempt at such an isolation would result in a category error. Therefore, the delineation between military and industry as particulars is but performative; in reality, their function is itself an interdependent and co-dependent relationship necessitating the need for stakeholders in and of themselves within this paradigm.

7.3 Formulating Design Requirements

Although various instruments and methods characteristic of the VSD methodology can be adopted to help designers distil and formalise the necessary requirements for any given design, the values hierarchy (i.e., Annex II, Figure 6; Chapter 6, Figure 2) is nonetheless useful as a way to illustrate and trace design requirements from norms to values, and vice versa. Figure 2 is one example of how to visualise the translation of higher-level values, from AI4SG norms through to technical design requirements (and bottom-up as well).

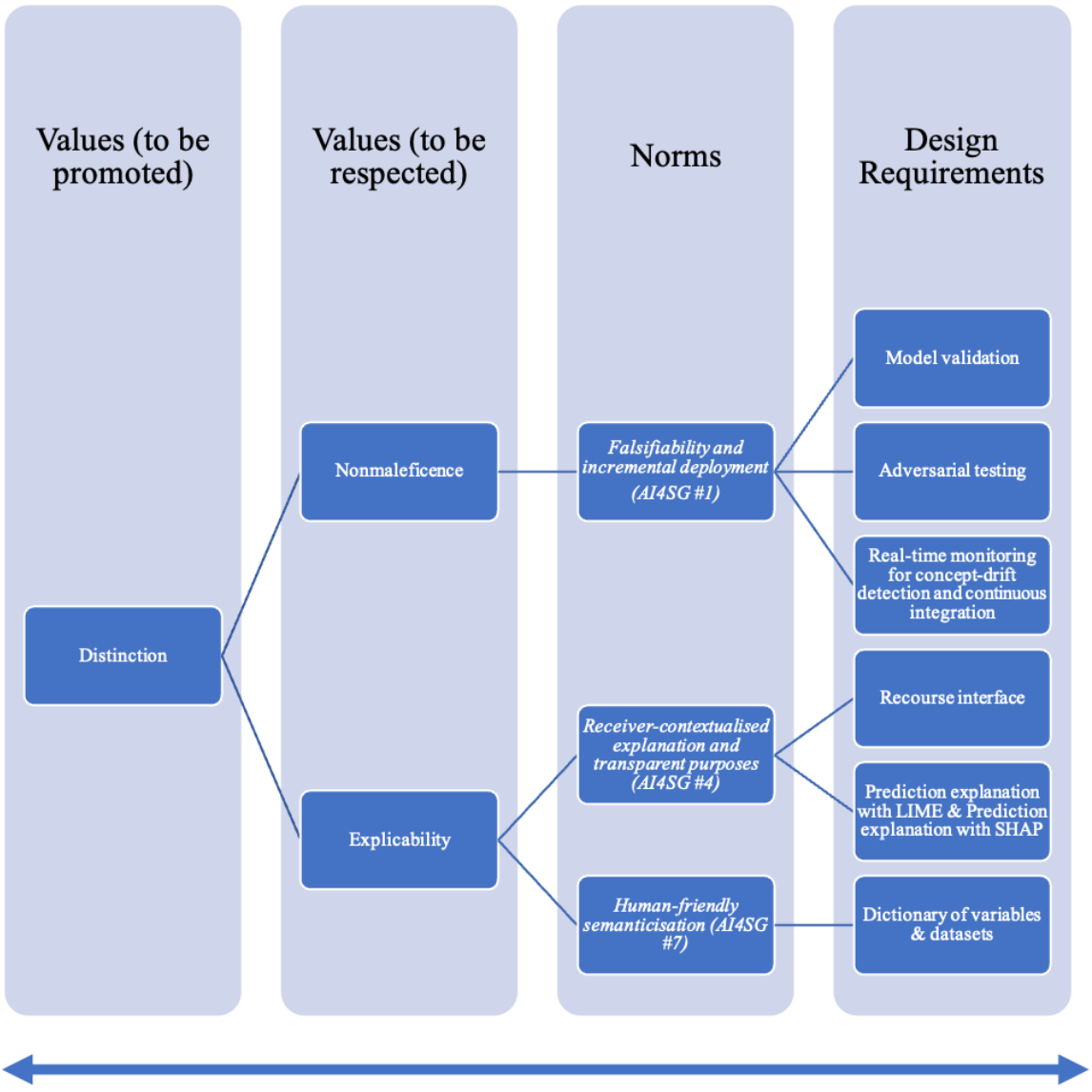

Figure 2. Bidirectional hierarchy of *distinction*, *explicability*, and *maleficence*

Here, *distinction* is chosen as the value to be promoted; it is then understood and/or satisfied as much as possible by the constraining values of *nonmaleficence* and *explicability* as the value to be respected. These two constraining values are then translated into AI4SG norms (1, 4, and 7, respectively), which in turn are transformed into design requirements. In this paradigm, AI4SG principles are adopted as *norms*, and rightly so, given that they are framed as imperatives by Floridi et al. (2020). Naturally, any given context of use, value, and specific technology will implicate any number of combinations, and there is no exclusive or exhaustive route to satisfy a value translation. It can move either in a bottom-up (or left to right, as Figure 2 illustrates; design requirements → norms → values) or a top-down (right to left; values → norms → design requirements) direction, as shown

above (cf., Longo et al., 2020). *Situational fairness* could just as easily, and probably should, be used as the normative tool for operationalising other values, including *explicability* (i.e., transparent dataset collection, use, storage, and destruction as well as the use of other methods such as predictive explanations and recourse interfaces; Yang et al., 2020) as well as *justice* (i.e., promoting non-discriminatory practices through unbiased compliance; e.g., using for Fairness Warnings and/or Fair-MAML examples described by Slack et al., 2020).

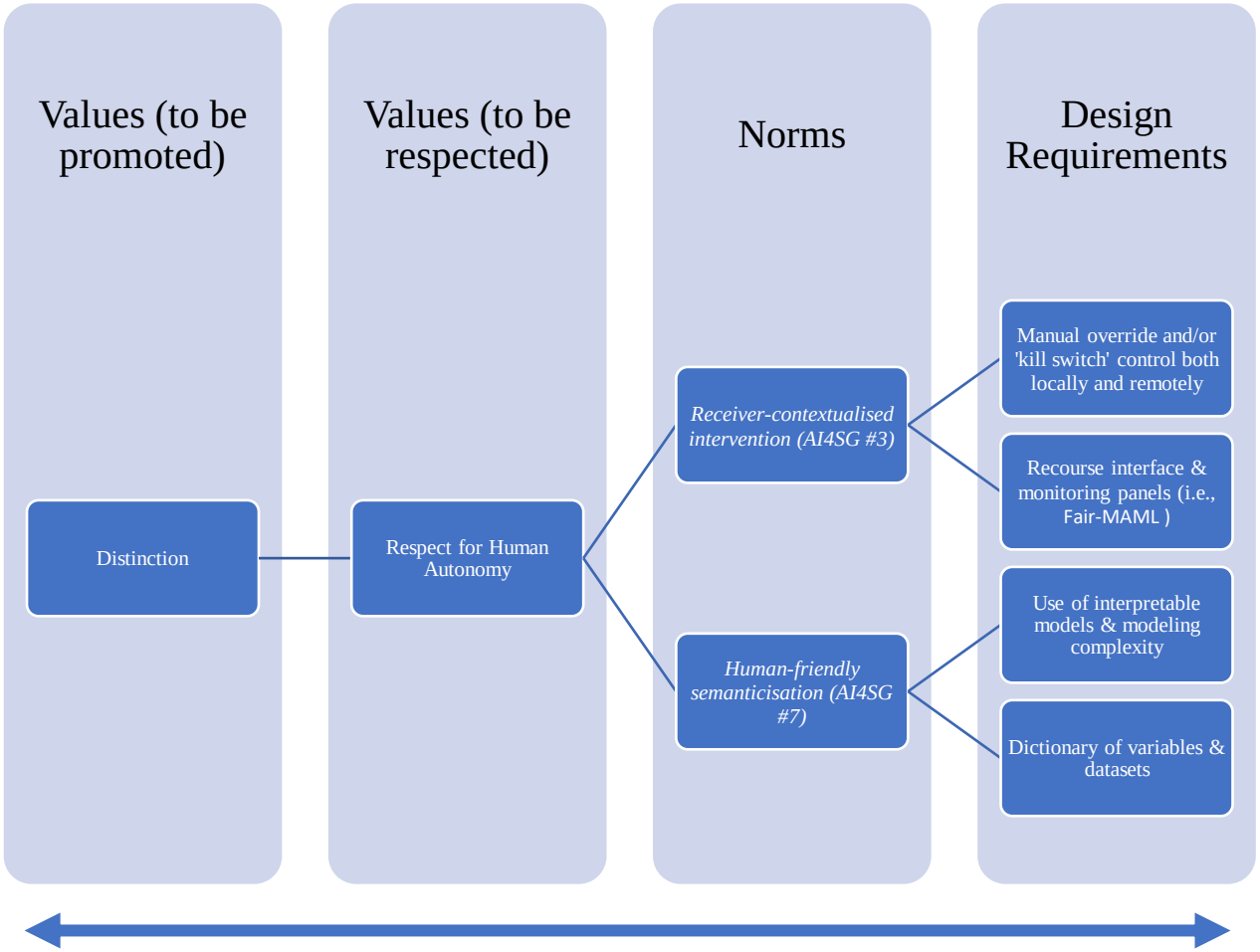

Figure 3. Bidirectional hierarchy of *distinction* and *respect for human autonomy*

Figure 3 again demonstrates that, as stated above, there is no exhaustive way of satisfying a value through its translation into design requirements. Here, the same higher value of *distinction* as the one selected for promotion can be understood in conjunction with *respect for human autonomy* – an organic coupling, since their definitional interrelation can be translated through AI4SG #3 and #7 (among which the latter is also fundamentally linked to the value of *explicability* in Figure 2,

effectively showcasing how the different values and AI4SGs are all interrelated and not ordered by rank). Human-friendly semantics are necessary here for the relevant moral agents in the design and the use of AWS chains to understand the nature and rationale of a specific action on the part of the system. They are also an epistemic necessity for proper receiver-contextualised intervention so as to balance the autonomy of both the system and the user based on the needs and capacities of both entities. Here, direct operational control is made possible (i.e., part of the *optionality* of AI4SG #3) while still permitting the system to retain full technical autonomy. Wirth regards to the dangers potentially posed by AWSs, full manual override can be permitted for such systems, in addition to recourse interfaces and/or monitoring panels that ensure full-lifecycle monitoring (e.g., Fair-MAML is one of many examples of a technical option for this norm).

At a functional level, the normative structure of the AI4SG norms prevents (most) ethical harms associated with AI systems. However, they do not guarantee per se that new AI applications will actively contribute to the greater social good. The higher-level values listed above (i.e., the EU HLEG AI), in conjunction with the related *real* operationalisation of the LOACs, allows for the development of more salient AI systems that contribute to global beneficence (i.e., international normative compliance). This multi-tiered approach of coupling AI specific and stakeholder values, along with their application to LOAC attainment via AI4SG norms, can mitigate dangers posed by the ethical white-washing that occurs through the legitimisation of AI technologies that do not respect certain fundamental ethical principles (Bietti, 2020; Metzinger, 2019; Reuters, 2021; Sloane, 2019).

Nonetheless, it is becoming increasingly apparent how designers can begin to design *for* MHC with regards to the *design level* thereof. In order to saliently design for maximal reason-responsiveness to both distal and proximal values, the design level should follow the requirements outlined in Chapters 5 and 6 on how to accurately embody values in sociotechnical systems. In the illustration above, for example, avoiding nonmaleficence can be translated into the norm of *falsifiability and incremental deployment* (AI4SG #1) as a means to attain one of the necessary ingredients for responsibly embodying values in AI systems; i.e., the ability for redesigning through continual lifecycle monitoring. In Figure 2, this can be incrementally attained through technical design requirements such as model validation, adversarial training, and real-time monitoring to ensure the continuity of the build-in concepts and allow for the possibility of a redesign otherwise. This real-time monitoring throughout the lifecycle of an AWS enables the accurate mapping of both higher-order values such as the LOACs to be continually validated in situ while also attending to value drift in the event of recalcitrance.

This is similarly true for the constraining value of *explicability*. Part of the tracking condition is the presence of at least one agent (individual or group) along the design and use chain of an AWS who understands the abilities and limitations of the system. As I argued earlier, decisions to employ such system, regardless of the cognitive clarity of the system itself, are also constrained by the *operational level* of MHC as a predicate of weaponeering choices. Nonetheless, this is necessary for the tracking condition to obtain; thus, even permitting higher levels of autonomy under this understanding of MHC, a system must be sufficiently explicable that such abilities and constraints are continuously accessible by the relevant moral agents in the design and use chain. Receiver-contextualised explanations and transparent purposes are the normative factors in which this concept can be most aptly understood. Translated into certain preliminary design requirements, this can be obtained via the implementation of recourse interfaces and predictive explanation tools to allow for more real-time understanding of system behaviour. Such tools are likewise interrelated with satisfying and/or strengthening the other values, such as nonmaleficence, in the above figure, given that more accurate, real-time, and transparent purposes as well as predictive explanation tools enable better lifecycle monitoring and more proximal redesigning to take place in the iterative process of VSD.

In any event, this type of visualisation can be used across different sources as listed above, such as the LOACs and stakeholder values, to determine how accurately related values can produce both similar and different technical design requirements. Future research projects can approach this empirically by taking any particular fully AWS variant and providing thorough value-design requirement translations to determine its effectiveness. All in all, the present goal is to more effectively design *for* various values in mind – ones that are often erroneously conflated if not completely side-lined.

7.4 Prototyping

In a recent Washington Post piece regarding the employment of AWSs as a means of more "ethical" warfare, one of the interviewees, William B. Roper Jr., a foreign policy strategist who served as the 13th Assistant Secretary of the Air Force for Acquisition, Technology, and Logistics and one of the Pentagon's chief proponents for the adoption of AI technologies, stated, "It doesn't make sense to study anything in the era of AI […]. It's better to let the AI start doing and learning, because it's a living, breathing system, very much like a human, just silicon-based" (Fryer-Biggs, 2021). This could not be further from reality, both in regards to the explicit push towards direct deployment without falsifiability tests and incremental deployment (i.e., AI4SG #1) and in terms of the erroneous

equivocation of human learning to that of machine learning. With regards to systems capable of such destruction, more prudence – not less – is needed in order to ensure that they remain lawful, and thus the adoption of a viable weaponeering option for military operations (i.e., under MHC). Part of this is a necessary orientation towards making falsifiability and incremental deployment – i.e., prototyping and full-lifecycle monitoring – a critical part of designing *for* MHC.

According to the design requirements laid out in the previous step, prototyping involves building mock-ups of the technology in question. This means that the technology is removed from the more controlled space of the laboratory or design space and built *in situ* – which, of course, implies direct and indirect stakeholder values. At this point, various design decisions may prove to be recalcitrant, or otherwise unforeseen recalcitrant behaviour emerges to involve other values. At this point, given the limited deployment of the technology, it can be recalled into the design space so that corrective modifications can be implemented. With regards to fully AWSs, the motivation behind their development is not as urgent as e.g., SARS-CoV-2 contact/tracing technologies, which are spurred on by the global crisis conditions and therefore resist slower prototyping and limited testing in favour of direct deployment. The development of fully AWSs, then, need not and should not follow the unwise route of direct deployment, given the significant risks that AI systems possess, particularly ones predicated on such large quantities of data with direct lethal capabilities. Small-scale deployment or in-house testing of the efficacy and fidelity of the ability of a system to be reason-responsive in potentially complex and dynamic scenarios – vis-à-vis war game scenarios – are a necessary (albeit insufficient) condition for the responsible development of an AI system of this type to ensure that it can aid in the achievement of positive ethical/societal values (i.e., beneficence, justice, explicability, autonomy, and the associated distal LOACs) while reducing the ethical (AI) risks (i.e., nonmaleficence).

It should be especially stressed that prototyping should not be restricted to testing the proper technical functioning of an app; it should also account for behavioural and societal effects as well as their ultimate impact on the values. The fully AWS is a case in point here. While some values – such as *explicability* and *respect for human autonomy* – can be designed within a system through technical choices, including complexity modelling, proper data collection and categorisation, recourse interfaces, and other auxiliary tools, some of the other concerns require insight into the behavioural effects of such AWSs. These behavioural effects are very difficult, if not impossible, to reliably predict without some form of prototyping or, at least, small-scale in-situ testing. It would therefore be advisable to conduct a number of trials for such systems that scale up through settings of increasing size, starting from very small-scale testing with mock-ups (not unlike what is done in medical experiments with new drugs).

Such testing trajectories might also reveal new values of significance that need to be considered, which can thereby trigger a new iteration of the development cycle.

7.5 Conclusions

The second part of this thesis discussed how AI systems can pose certain challenges for the VSD approach to technology. These challenges primarily result from the use of ML approaches to AI, the approaches that are most probable to be adopted for AWS design. Machine learning poses two challenges to VSD. First, it may be opaque (to humans) how an AI system has learned certain things, which requires attention to such values as *transparency*, *explicability*, and *accountability*. Second, ML may lead AI systems to adapt themselves such that they "disembody" the values that have been embodied in them by VSD designers. In order to deal with these challenges, an extension of VSD to the whole lifecycle of AI systems design was proposed. More specifically, I discussed how the AI4SG principles proposed by Floridi et al. can be integrated as norms in VSD when considering AI design. In order to integrate the AI4SG principles into a more systematic VSD approach, I presented a design process that consists of four basic iterative steps: contextual analysis, value identification, translation of values into design requirements, and prototyping. At the core of this model is a two-tiered approach to values in AI consisting of (1) a real commitment to contributing to beneficence via an explicit design orientation *for* the LOACs through AI and (2) the formulation and strict adoption of a number of concrete AI4SG norms. Without the first tier, AI4SG factors may help to prevent (most) categories of ethical harm, but there is absolutely no guarantee that new AI applications will actively contribute to the greater social good. Meanwhile, the second tier eliminates the risk of societal challenges and of LOACs being used to legitimise AI technologies that do not respect certain fundamental ethical principles; i.e. the danger of ethical white-washing (which is already visible on the webpages of some large companies). In addition, it is important to pay attention to contextual values – or at least to the contextual interpretation of the values from the two tiers. This is necessary for understanding why certain values are at stake for a specific application and how to translate the relevant values into design requirements.

References

Arkin, R. C. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture, part I: Motivation and philosophy. *Proceedings of the 3rd International Conference on Human Robot Interaction (HRI 2008)*, 121. https://doi.org/10.1145/1349822.1349839

Bietti, E. (2020). From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 210–219.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fryer-Biggs, Z. (2021, February 17). *Future warfare will feature autonomous weaponry.* The Washington Post. https://www.washingtonpost.com/magazine/2021/02/17/pentagon-funds-killer-robots-but-ethics-are-under-debate/?arc404=true

Gabriel Wood, N. (2020). The problem with killer robots. *Journal of Military Ethics*, *19*(3), 220–240. https://doi.org/10.1080/15027570.2020.1849966

Guetlein, M. A. (2005). *Lethal autonomous weapons – Ethical and doctrinal implications*. Joint Military Operations Department, U.S. Naval War College.

ICRC. (2002). *Introduction to the law of armed conflict – Basic knowledge*. www.icrc.org

Longo, F., Padovano, A., & Umbrello, S. (2020). Value-oriented and ethical technology engineering in industry 5.0: A human-centric perspective for the design of the factory of the future. *Applied Sciences*, *10*(12), 1–25. https://doi.org/10.3390/APP10124182

Mecacci, G., & de Sio, F. S. (2019). Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology*, *22*, 1–13.

Metzinger, T. (2019). *Ethical washing machines made in Europe*. Tagesspiegel. https://background.tagesspiegel.de/ethik-waschmaschinen-made-in-europe

Pasquale, F. (2017). Toward a fourth law of robotics: Preserving attribution, responsibility, and explainability in an algorithmic society. *Ohio State Law Journal*, *78*, 1243.

Reuters. (2021, January 26). *US has "moral imperative" to develop AI weapons, says panel | Weapons technology*. The Guardian. https://www.theguardian.com/science/2021/jan/26/us-has-moral-imperative-to-develop-ai-weapons-says-panel

Slack, D., Friedler, S. A., & Givental, E. (2020). Fairness warnings and Fair-MAML: Learning fairly with minimal data. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 200–209. https://doi.org/10.1145/3351095.3372839

Sloane, M. (2019). Inequality is the name of the game: Thoughts on the emerging field of technology, ethics and social justice. *Proceedings of the Weizenbaum Conference 2019 "Challenges of Digital Inequality – Digital Education, Digital Work, Digital Life,"* 9.

Umbrello, S., Capasso, M., Balistreri, M., Pirni, A., & Merenda, F. (2021). *Expanding care-centered value sensitive design: Design care robots with AI for social good norms*.

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

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 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

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., & Russakovsky, O. (2020). Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 547–558. https://doi.org/10.1145/3351095.3375709

8 Conclusion

Considering the ongoing international discussions on the ethics and legality of AWSs and whether we can have MHC over certain levels of autonomy, this thesis focuses on a more pragmatic understanding of autonomy in the military domain as well as how to design *for* MHC in ensuring the responsible design and deployment of certain types of fully AWSs. To this end, this thesis focuses on the question: *how can we understand autonomy so as to construct a more grounded conception of MHC, and how can we explicitly design for MHC?* In response to this question, Chapter 2 argued that we cannot jump right into questions of autonomy or MHC without first acknowledging the sociotechnicity of artefacts and, by extension, that a systems thinking ontology provides a solid framework for conceptualising the interconnectedness between AWSs and their social contexts (e.g., military, industry, legal norms, and human agents, among others). Avoiding this ontological step may cause us to miss the forest for the trees in correctly understanding MHC. In fact, systems thinking and systems engineering originated in the military domain for this very reason; viz., the complex networks of agents and technologies make salient design and deployment difficult, if not impossible, without being viewed and designed through such lenses and approaches.

Aligned with the complexity of the various sociotechnical contexts in which we need to understand the design and use of fully AWSs, Chapter 3 outlined two different levels of abstraction (LoAs) that, when coupled, form a more holistic understanding of MHC. More precisely, the MHC of AWSs cannot be divorced from *actual* military operations planning and management. This managerial and planning abstraction is essential to a systems thinking approach (and thus systems engineering) and how autonomy is supported and constrained *per se*, regardless of whether the agent is a human soldier or an AWS. By the very nature of the military system, AWSs can be said to be under a certain level of MHC, all else being equal on the technical side (i.e., it is fully responsive to all the intentions and expectations of military leaders). I call this the *operational level* of control in understanding MHC. However, this is barring the actual technical design, which also needs to align with the LOACs and military priorities. The *design level* of control is a function of the *reason-responsiveness* of an AWS to the moral reasons of the relevant agents in the design history and use chain of a system. Once again, the systemic nature of AWSs and their context plays a crucial role here. The level of design does not place the burden of control exclusively on the end-user but on at least one of the moral agents in this design/use chain. This does not necessarily mean a single human agent; it can also include

supraindividual agents such as the military itself and naturally its partnerships with the industrial firms that design and construct AWSs.

Chapter 4 then united these two levels of abstraction into what is arguably a more comprehensive understanding of MHC for AWSs. It did so by placing AWSs within their context of use (i.e., the operational level), which *de facto* supports and constrains specific uses of these types of system as well as all agents and systems within that domain. The chapter also examined the level of design and the technical functionality of an autonomous system and its sensitivity to the moral reason(s) of the relevant agent(s). If a system is designed to be maximally responsive to these types of reasons, then it cannot only be designated under MHC; autonomy *per se* is not *mala in se* as many ban proponents believe, but can actually be used to augment MHC. Hence, the marriage of both levels of MHC (i.e., operational and design) was shown to be symbiotic with regards to MHC. Here, the argument is that military operations *typically already* constrain the autonomy of any and all agents within the military-industrial complex as a function of the procedures that necessarily take place *a priori* to the deployment of force (i.e., the operational level). Likewise, the close cooperation between institutions and infrastructures that constitute the military-industrial complex (e.g., the military, industry, government, and legislative norms, among others) form the *supraindividual* agent that can be called the possessor of MHC if the design history can be *traced* and its behaviours can be *tracked* to the relevant moral agents (i.e., the MIC). These two levels of abstraction warrant closer cooperation within the MIC to enable more accurate mapping of the moral intentions of the relevant agents onto the AWSs that are being developed/deployed. The consequence here is that if MHC obtains across both levels of control, then autonomy *per se* is not the problematic vector; it can actually be increased, thereby increasing MHC.

Value-sensitive design has been adopted as a principled approach for the design of various existent as well as futuristic/transformative technologies. The VSD approach is fundamentally predicated on the interactional stance on technology – or, more precisely, that societal and social factors co-construct and co-vary with technological artefacts. Part of this approach is that technologies embody values. However, AI systems, like projected (fully)AWS, that employ machine learning (ML) and/or artificial neural networks are often opaque, and the values that they may (dis)embody can therefore be unforeseen or unforeseeable. Chapter 5 discussed the myriad ways in which technologies embody values and how they fit within the larger systems thinking approach, as well as how to more saliently frame the embodiment of values *for* AI systems such as AWSs. Because ML systems (often) learn in ways that are opaque to humans, we need to pay close attention to values such as transparency,

explicability, and accountability. To address this issue, as well as the potential "disembodiment" of certain values over time, Chapter 6 proposed a threefold modified VSD approach: (1) integrating a known set of VSD principles (AI4SG) as design norms, from which more specific design requirements can be derived; (2) distinguishing between values that are promoted and respected by the design to ensure outcomes that cause no disproportionate harm as well as actively promote just war; and (3) extending the VSD process to encompass the whole lifecycle of an AI technology, so as to monitor unintended value consequences and facilitate redesigning as needed.

Finally, in Chapter 7, I demonstrated the AI4SG-VSD approach described in the two preceding chapters with the AWS as the use case. In doing so, I outlined the various values to be promoted (i.e., the LOACs), the constraining values to be respected (i.e., the EU HLEG AI), as well as the AI4SG norms as a means for translating these abstract values into technical design requirements. The hierarchy of values was chosen as the tool for illustrating how designers can begin to conceptualise this transition into designing *for* values rather than doing so *ex post facto*, ad hoc, or not at all. Furthermore, the chapter discussed how full-lifecycle monitoring and incremental deployment into an envelope of safe can be used to determine the emergent behaviours and consequent implicated values, which in turn informs us if a redesign is necessary for a system. In the event that this cannot be accomplished, then such types of systems should be considered *de facto*, or otherwise prohibited on account of the associated risks of bypassing such an approach.

Thus, *is meaningful human control over fully AWSs possible?* I have aimed to argue yes, but only in some instances. This thesis makes a strong case for a nuanced answer; more specifically, as things stand, this thesis argues for the possibility of achieving not only MHC, but MHC with greater autonomy in certain aerial fully AWSs. Through strong partnerships between and within the MIC, clearer design histories and use chains can be determined so as to communicate and transfer knowledge between stakeholders adequately. Similarly, the salient and lawful weaponeering of systems such as AWSs is directly contingent on these types of cooperation, and it bridges any epistemic gaps that designers and users may have. Therefore, the systems have to be designed such that they are sufficiently reason-responsive to these supraindividual agents. If what I have tried to demonstrate in this thesis obtains, then this type of MHC is not only possible but also preferable, and it should be seriously considered by both armed forces and policymakers as a salient middle path that can satisfy military intentions as well as restrict more egregious forms of AWSs. In either case, one may argue that the paths laid out in this thesis are overly constraining. The objectives of human-like fully AWSs, ground-based or otherwise, are excluded in this landscape, as they arguably should be. To reiterate a

similar thought by Scott Robbins, “the future of AI is not, and should not be, machines from which human moral responsibility has been removed, but in machines that enhance our ability to be morally responsible” (Robbins, 2020, p. 172). This is precisely what makes MHC meaningful.

As mentioned throughout, this thesis takes on one very particular, but central argument often proposed by ban proponents as a cogent reason for the prohibition of fully AWS. That being said, there are other strong arguments for why certain types of AWS should be prohibited; the argument proposed here does not necessary speak to the arguments proposed for those positions. There are arguments that AWS will have deleterious effects on human dignity, something that I personally am partial to given that I hold central many of martial values that are side-lined when war is abdicated to technical systems rather than humans. However, dignity, despite the common heuristic understanding of the term, is nonetheless hard to narrow down in such a way that is minimally sufficient for binding international treaties. Likewise, I have argued elsewhere that certain aerial (fully)AWS may actually be permissible under certain laws of armed conflict, in particular *hors de combat* (combat status). Although this is not the place to go into depth with this argument, what that work betrays is the need for systems that are capable of highly-contextualised situational awareness, something that is currently technically limited. As a consequence, until AWS are capable of such gradual contextual awareness they risk contravening the laws of armed conflict and thus should be prohibited pursuant of such.

Regardless, as the contexts of war continue to change as part and parcel of the technological development that characterises sociotechnical systems, so too will the specifics proposed in this thesis. In particular, the design requirements in the final chapter will certainly not survive the test of time, as new and potentially better technical mechanisms emerge to satisfy the norms (which may also change) and values of responsible design. What is certain, however, is that such values and norms will remain interconnected with the technology in question. Similarly, designers will always need a principled approach to actually design *for* human values rather than treating them as mere afterthoughts. At the very least, this thesis aims to support the latter two, whereas the former will require constant updating, knowledge transfer, and closer partnerships to ensure that MHC always obtains.

In summary, this thesis aims to argue more broadly that we should resist the totalising narrative that equates the autonomy of AWS in any way to that of human agents, military or otherwise. Autonomy as described here is not *mala in se*; it can actually augment MHC rather than diminish it *mutatis mutandis*. As such, Terminator-derived narratives must be resisted if we wish to achieve any semblance of responsible innovation, even that mired by the fog of war. Doing otherwise risks the detriment of not only a lack of such responsibility, but any MHC whatsoever. The middle path

provided here offers a nuanced way of understanding MHC in addition to *autonomy* as the vector of discussion for AWSs.

Summary

Artificial intelligence – specifically machine learning (ML) – is becoming ever more ubiquitous in society, in terms of its presence in both everyday technologies such as cell phones and advanced applications such as fast trading algorithms in the stock market. Coupled with the almost unfathomable quantity of data that characterised big data, AI systems have become ever more nebulous, opaque, and difficult – if not impossible – to understand. Their ability to process inhuman quantities of data and make decisions in an increasingly autonomous fashion have enabled them to make their mark not only in civilian spheres such as research laboratories, pharmaceuticals, and stock markets, but also naturally in the military domain.

Once a domain limited to humanity (and of course the animal kingdom as well), warfare is marked by races towards greater supremacy and technological prowess in a bid to ensure victory and maintain hegemony. Artificial intelligence systems were arguably born in the military domain, and they continue to be a driving technology powering many current military systems in most arsenals, especially global superpowers such as the United States, the United Kingdom, Russia, and China. In particular, the growing global debate on the use of AI-powered autonomous weapons has merited attention. The increasing abdication of human control and executive authority to machines becomes worrisome when we envision the natural consequence of this trend towards total abdication and thus full autonomy in selecting and engaging targets without human intervention or consent. These fully autonomous weapons systems (AWSs) are the central object of study in this thesis. More specifically, this thesis aims to explore how to ensure that meaningful human control (MHC) over AWSs is designed *for*. It argues that, in order to have a sufficiently comprehensive understanding of MHC, such a theory must account for the real military procedures of how operations are conducted, as well as the design histories and relevant moral agents within the military-industrial complex, which are also fundamental to the design and deployment of such sociotechnical systems.

This thesis is functionally separated into two distinct parts. In Part I (consisting of Chapters 1–4), a systems theory ontology is proposed as a unifying substratum for understanding MHC as well as how it can be designed in the proceeding part. Here, the focus is on two different levels of abstraction regarding MHC: operational and design. They are posited as both necessary, at least in terms of AWSs, for sufficient MHC to be attained. The nuance that surfaces here is that full autonomy per se is not necessarily as problematic as some detractors of AWSs have claimed, and that, in some cases, greater autonomy can augment MHC as a function of proper design. Part II focuses on the value-sensitive design (VSD) approach as a methodology. (1) It exists upon the same systems ontology foundation

proposed in Part I, and (2) *mutatis mutandis* VSD forms a sufficient – at least preliminary – approach to designing *for* MHC across both levels of abstraction.

In order to bridge the levels of abstraction and thereby conceptualise a unified theory of MHC over AWSs, as well as to subsequently unify this conception of MHC with a design approach that is capable of designing *for* it (i.e., VSD), chapter 2 proposes systems thinking as the ontological substrata. The main reason for adopting this approach is that it (implicitly) characterises the two levels of abstraction for understanding MHC. The operational level of control is characterised by a plurality of actors and networks that complicates but also constitutes how military operations are structured, planned, and conducted. Likewise, the design level of control is fundamentally built on the notion of tracking and tracing networks of systems and actors within both the use and the design histories of those systems. In addition, systems thinking is the theoretical framework from which systems engineering derives. It is essentially the practical and managerial implementation of a systems thinking ontology, whereas VSD exists as a sort of parallel approach to the systems thinking design methodology

To couple the various levels of abstraction, chapter 3 builds on the literature review of Annex I, in which both Ekelhof and Santoni de Sio's works on MHC, among others, are explained. In this chapter, the approaches presented in these papers are discussed, in addition to how we can begin to view those approaches as symbiotic in terms of their systems thinking affinities. The initial groundwork is then laid for understanding how they both complement each other without encumberment.

The marriage of both levels of MHC (i.e., the operational and design levels) is demonstrated to be symbiotic with regards to MHC. Here, in chapter 4, the argument is that military operations *always already* constrain the autonomy of any and all agents within the military-industrial complex as a function of the procedures that necessarily take place *a priori* to the deployment of force (i.e., the operational level). Close cooperation between institutions and infrastructures that constitute the military-industrial complex (e.g., the military, industry, government, and legislative norms) likewise form the *supraindividual* agent that can be said to be the possessor of MHC, if the design history can be *traced* and its behaviors can be *tracked* to the relevant moral agents (i.e., MIC). These two levels of abstraction warrant closer cooperation within the MIC so as to allow more accurate mapping of the moral intentions of the aforementioned agents onto AWSs that are being developed/deployed. The consequence here is that, if MHC obtains across both levels of control, then not only is autonomy *per se* not the problematic vector, but it can actually be increased, thereby increasing MHC.

Value-sensitive design has been adopted as a principled approach to designing various existent as well as futuristic/transformative technologies. The VSD approach is fundamentally predicated on the interactional stance towards technology – or, more precisely, that societal and social factors co-construct and co-vary with technological artifacts. Part of the rationale behind this approach is that technologies embody values. However, AI systems that employ machine learning (ML) and/or artificial neural networks are often opaque, and thus the values that they may (dis)embody can be unforeseen or unforeseeable. Chapter 5 discusses the different ways in which technologies embody values and how they fit within the larger systems thinking approach, as well as how to more saliently frame the embodiment of values *for* AI systems such as AWSs.

As ML systems (often) learn in ways that are opaque to humans, we need to pay attention to values such as transparency, explicability, and accountability. To address this issue, as well as the potential "disembodiment" of certain values over time, chapter 6 proposes a threefold, modified VSD approach: (1) integrating a known set of VSD principles (AI4SG) as design norms, from which more specific requirements can be derived; (2) distinguishing between values that are promoted and respected by the design to ensure outcomes that not only prevent disproportionate harm but also actively promote just war; and (3) extending the VSD process to encompass the whole lifecycle of an AI technology, so as to monitor unintended value consequences and redesign as needed.

The AI4SG-VSD approach described in the previous two chapters is employed with the AWS as the use case. In doing so, chapter 7 outlines the values to be promoted as much as possible (e.g., the LOACs), the (constraining) values to be respected as much as possible (e.g., the EU HLEG AI), as well as the AI4SG norms as a means for translating these abstract values into technical design requirements. The value hierarchy is chosen as the tool for illustrating how designers can begin to conceptualise this translation to design *for* values rather than *ex post facto*, ad hoc, or not at all. Likewise, I discuss how full-lifecycle monitoring and incremental deployment into an envelope of safe use to determine the emergent behaviours and consequent implicated values can be used to evaluate whether a system requires a redesign. In the event that this cannot be done, such types of systems should be considered *de facto*, or otherwise prohibited, given the associated risks of bypassing such an approach.

If we turn back to the beginning and inquire whether MHC is possible for AWSs, the answer is yes, but not without some caveats. First, for the MHC of fully AWSs to obtain, MHC must couple two levels of abstraction: operational and design. In doing so, *prima facie* at least certain forms of fully AWSs are permitted (e.g., aerial fully AWSs). This is not a banal or trivial point; aerial warfare, both that conducted for the purposes of superiority and air strikes, is becoming an increasingly preferable

option, and the recent trend towards it merits more attention. This trend, mainly on account that aerial warfare capabilities are a force multiplier, means that fully AWSs are most likely to appear in this vector, rather than the more problematic, ground-based imaginings of terminator-type AWSs. In any case, if this point holds water, then full autonomy is not *mala in se*, and it is rendered unproblematic, at least in certain types of AWSs.

Riassunto

L'intelligenza artificiale, in particolare il Machine Learning (ML), sta diventando sempre più onnipresente nella società, sia nelle tecnologie quotidiane come i cellulari, e sia nelle applicazioni avanzate come gli algoritmi di trading veloce nel mercato azionario. Insieme alla quantità quasi insondabile di dati che caratterizzano i big data, i sistemi di intelligenza artificiale sono diventati sempre più nebulosi, opachi e difficili, se non impossibili, da capire. La loro capacità di elaborare quantità disumane di dati e di prendere decisioni in modo sempre più autonomo ha permesso loro di lasciare il segno non solo nelle sfere civili come laboratori di ricerca, prodotti farmaceutici e mercati azionari, ma anche naturalmente nel dominio militare**.**

Dominio che una volta era limitato all'umanità (e ovviamente anche al regno animale), che ha portato la guerra a gare verso una maggiore supremazia e abilità tecnologiche nel tentativo di garantire la vittoria e mantenere l'egemonia. I sistemi di intelligenza artificiale sono probabilmente nati nel dominio militare e continuano ad essere una tecnologia trainante che alimenta molti attuali sistemi militari nella maggior parte degli arsenali, in particolare superpotenze globali come Stati Uniti, Regno Unito, Russia e Cina. In particolare, ha meritato attenzione il crescente dibattito globale sull'uso di armi autonome alimentate dall'intelligenza artificiale. La crescente abdicazione del controllo umano e dell'autorità esecutiva alle macchine diventa preoccupante quando si immagina la naturale conseguenza di questa tendenza all'abdicazione totale e quindi alla piena autonomia nella selezione e nel coinvolgimento di obiettivi senza intervento o consenso umano. Questi sistemi d'arma completamente autonomi, *autonomous weapons systems* (AWS), sono l'oggetto centrale di studio in questa tesi. Più specificatamente, questa tesi mira ad analizzare la possibilità di progettare AWS per garantire il controllo umano significativo, *meaningful human control* (MHC). Sostiene che, al fine di avere una comprensione sufficientemente completa di MHC, una tale teoria deve tenere conto delle reali procedure militari, ovvero di come vengono condotte le operazioni, l'avvenimento storico delle decisioni progettuali sulla tecnologia e gli agenti morali relativi all'interno del complesso militare-industriale, che sono fondamentali anche per la progettazione e la diffusione di tali sistemi sociotecnici.

Questa tesi è suddivisa in due parti distinte. Nella Parte I (composta dai capitoli 1–4), viene trattata un'ontologia della teoria dei sistemi, proposta come substrato unificante per comprendere MHC e come può essere progettato nella parte precedente. L'attenzione si concentra su due diversi livelli di astrazione di MHC: operativo e di design. Sono ritenuti entrambi necessari, almeno in termini di AWS, per ottenere un MHC sufficiente. La sfumatura che emerge è che la piena autonomia di per sé non è necessariamente così problematica come hanno affermato alcuni detrattori di AWS e che, in alcuni

casi, una maggiore autonomia può aumentare MHC in funzione di una corretta progettazione. La parte II si concentra sull'approccio di progettazione sensibile al valore (VSD) come metodologia. (1) Il quale esiste sulla stessa base di ontologia dei sistemi proposta nella Parte I, e (2) "*mutatis mutandis*" VSD costituisce un approccio sufficiente - almeno preliminare - alla progettazione per MHC attraverso entrambi i livelli di astrazione.

Al fine di colmare i livelli di astrazione e quindi concettualizzare una teoria unificata di MHC su AWS, nonché di unificare successivamente questa concezione di MHC con un approccio progettuale in grado di progettare per esso (cioè, VSD), il capitolo 2 propone il pensiero sistemistico(systems thinking) come substrato ontologico. La ragione principale per adottare questo approccio è che (implicitamente) caratterizza i due livelli di astrazione per la comprensione dell'MHC. Il livello operativo di controllo è caratterizzato da una pluralità di attori e reti i quali lo complicano maggiormente, ma costituisce anche il modo in cui le operazioni militari sono strutturate, pianificate e condotte. Allo stesso modo, discuto di come il monitoraggio dell'intero ciclo di vita e l'implementazione graduale in una zona di sicurezza viene utilizzata per determinare i comportamenti emergenti, e i conseguenti valori implicati i quali possono essere utilizzati per valutare se un sistema richiede una riprogettazione. Inoltre, il pensiero sistemico è il quadro teorico da cui deriva l'ingegneria dei sistemi, ed è essenzialmente l'implementazione pratica e gestionale di un'ontologia del pensiero sistemico, pertanto VSD esiste come una sorta di approccio parallelo alla metodologia di progettazione del pensiero sistemico.

Per congiungere i vari livelli di astrazione, il capitolo 3 si basa sulla revisione della lettura dell'allegato I, in cui vengono spiegati, oltre ad altro, i lavori di Ekelhof e Santoni de Sio sull'MHC. Vengono inoltre discussi gli approcci presentati in questi articoli, e come possiamo iniziare a vederli simbiotici in termini di affinità di pensiero sistemico. Sono quindi gettate le basi per capire come si completano a vicenda senza complicazioni.

Il matrimonio di entrambi i livelli di MHC (quello operativo e quello di progettazione) si è dimostrato dunque simbiotico. Nel capitolo 4 l'argomento è incentrato sul fatto che le operazioni militari già vincolano l'autonomia di tutti gli agenti all'interno del complesso militare-industriale in funzione delle procedure, che necessariamente si svolgono a priori del dispiegamento della forza (cioè, il livello operativo). La stretta collaborazione tra istituzioni e infrastrutture, le quali costituiscono il complesso militare-industriale (ad esempio, le norme militari, industriali, governative e legislative) forma allo stesso modo l'agente *sovraindividuale* , il quale si può dire essere il possessore di MHC soltanto se la storia del design può essere *tracciata* ei suoi comportamenti possono essere *rintracciati*

dagli agenti morali rilevanti (cioè, MIC). Questi due livelli di astrazione garantiscono una più stretta cooperazione all'interno del MIC, in modo da consentire una mappatura più accurata delle intenzioni morali dei suddetti agenti sulle AWS che vengono sviluppate / distribuite. La conseguenza qui è che, se l'MHC si ottiene attraverso entrambi i livelli di controllo, non solo l'autonomia di *per sé* non è il vettore problematico, ma può effettivamente essere aumentata, aumentando così l'MHC.

Il Value-sensitive design è stato adottato come approccio di principio per la progettazione di varie tecnologie esistenti e futuristiche/trasformative. L'approccio VSD è fondamentalmente basato sulla posizione interazionale nei confronti della tecnologia o, più precisamente, che il sociale e i fattori sociali co-costruiscono e co-variano con gli artefatti tecnologici. Parte della logica alla base di questo approccio è che le tecnologie incarnano i valori. Tuttavia, i sistemi di intelligenza artificiale che impiegano l'apprendimento automatico (ML) e/o le reti neurali artificiali sono spesso opachi e quindi i valori che possono (dis)impersonare possono essere imprevisti o imprevedibili. Il Capitolo 5 discute i diversi modi in cui le tecnologie impersonano i valori e come si adattano all'approccio sistemico più ampio, nonché come inquadrare in modo più saliente l'incarnazione dei valori *per* i sistemi di intelligenza artificiale come gli AWS.

Poiché i sistemi di machine learning (spesso) imparano in modi che sono opachi per gli esseri umani, dobbiamo prestare attenzione a valori come trasparenza, spiegabilità e responsabilità. Per affrontare questo problema, così come la potenziale "disincarnazione" di determinati valori nel tempo, il capitolo 6 propone un triplice approccio VSD modificato: (1) integrando un noto insieme di principi VSD (AI4SG) come norme di progettazione, da cui possono essere derivati requisiti più specifici; (2) distinguendo i valori promossi e rispettati dal progetto per garantire risultati che non solo prevengano danni sproporzionati, ma promuovano anche attivamente la guerra giusta; e (3) estendere il processo VSD per comprendere l'intero ciclo di vita di una tecnologia di intelligenza artificiale, in modo da monitorare le conseguenze del valore non intenzionale e se necessario riprogettarla.

L'approccio AI4SG-VSD descritto nei due capitoli precedenti viene utilizzato con AWS come caso d'uso. Pertanto, il capitolo 7 delinea i valori da promuovere il più possibile (ad esempio, i LOAC), e quelli (vincolanti) da rispettare il più possibile (ad esempio, EU HLEG AI), nonché le norme AI4SG come mezzo per tradurre questi valori astratti in requisiti di progettazione tecnica. La gerarchia dei valori viene scelta come strumento per illustrare come i designer possono iniziare a concettualizzare questa traduzione per progettare *per valori* piuttosto che *ex post facto, ad hoc* o per niente. Allo stesso modo, discuto di come il monitoraggio dell'intero ciclo di vita e l'implementazione incrementale in un involucro di utilizzo sicuro per determinare i comportamenti emergenti ei conseguenti valori implicati

possano essere utilizzati per valutare se un sistema richiede una riprogettazione. Nel caso in cui ciò non possa essere fatto, tali tipi di sistemi dovrebbero essere considerati *de facto*, o altrimenti vietati, dati i rischi associati di aggirare tale approccio.

Se torniamo all'inizio e ci chiediamo se l'MHC è possibile per AWS, la risposta è sì, ma non senza alcuni avvertimenti. In primo luogo, per ottenere completamente l'MHC di AWS, quest'ultimo deve congiungere due livelli di astrazione: operativo e progettazione. In tal modo, *prima facie* sono consentite interamente solo alcune forme di AWS (ad esempio AWS completamente aeree). Questa non è un'osservazione banale; la guerra aerea, sia quella condotta a fini di superiorità e sia quella di attacchi aerei, sta diventando un'opzione sempre più preferibile, e la recente tendenza verso di essa merita maggiore attenzione A causa del fatto che le capacità della guerra aerea sono un indice di moltiplicatore di forza, è molto più probabile che in questo vettore appaiono gli AWS completi, al posto dell'immaginazione terrestre degli AWS terminator.

In ogni caso, se questo punto è valido, allora la piena autonomia non è *mala in se*, ed è resa non problematica, almeno in alcuni tipi di AWS.

ABOUT THE AUTHOR

Steven Umbrello (1993) obtained his Honours Bachelor of Arts (H.B.A) from the University of Toronto with a major in Philosophy and two minors, one in the Philosophy of Science and the other in Classical Civilizations. He went on then to obtain a Master of Arts (MA) from York University (Canada) in Science and Technology Studies while concurrently earning a Master of Science (MSc) in Epistemology, Ethics and Mind from the University of Edinburgh. During this seven year period Steven volunteered with numerous research groups and think-tanks, most notably the Global Catastrophic Risk Institute [GCRI] (2013-2018) and the Institute for Ethics and Emerging Technologies [IEET] (2015-). During his time at GCRI, and under the supervision of its Executive Director Seth Baum, Steven was guided on proper research practices and publishing ethics, leading him to publishing his first academic articles. Similarly, during his tenure at IEET, first as Assistant Managing Director (2015-2016) and then as Managing Director (2016-), he went on to continue his research and publishing in academic journals of repute that have culminated in this opus.