Comprehensive Human Oversight over Autonomous Weapon Systems

Verdiesen, E.P.

DOI
10.4233/uuid:5d444c43-0e3a-4912-838c-5a9c20ffee97

Publication date
2024

Document Version
Final published version

Citation (APA)
Verdiesen, E. P. (2024). *Comprehensive Human Oversight over Autonomous Weapon Systems*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:5d444c43-0e3a-4912-838c-5a9c20ffee97

Comprehensive Human Oversight over Autonomous Weapon Systems

Ilse Verdiesen

COMPREHENSIVE HUMAN OVERSIGHT OVER AUTONOMOUS WEAPON SYSTEMS

Ilse Verdiesen

Cover concept: Ilse Verdiesen
Cover: Ilse Modder (www.ilsemodder.nl)
Layout: Ilse Modder (www.ilsemodder.nl)
Print: Gildeprint Drukkerijen (www.gildeprint.nl)

ISBN:978-94-6496-075-4

Comprehensive Human Oversight over Autonomous Weapon Systems

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof.dr.ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Wednesday 3 April 2024 at 15:00 o'clock

By

Elizabeth Paulina VERDIESEN
Master of Science in Systems Engineering, Policy Analysis and Management,
Delft University of Technology, the Netherlands
born in Rotterdam, the Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Dr. M.V. Dignum	Delft University of Technology, promotor
Dr. F. Santoni de Sio	Delft University of Technology, promotor

Independent members:

Prof.dr. C. Jonker	Delft University of Technology
Prof.dr.ir L.M.M. Royakkers	Eindhoven University of Technology
Prof.dr. F. Osinga	Leiden University
A. Kaspersen MSc.	Carnegie Council for Ethics in International Affairs
Prof.dr.ir. M.F.W.H.A. Janssen	Delft University of Technology, reserve member

This dissertation and research are based on my own work that I conducted independently over the past six years. I have written all my published articles with co-authors which is common practise in my research field. At the start of each chapter, I clearly state in which articles parts of the chapter have been published.

I have written this dissertation in plural using the direct verb '**we**' as is customary in my discipline in recognition of all the comments and supervision of my promotors and fellow researchers to show that I greatly appreciate their support and mentoring.

For I firmly believe that the following applies both in the scientific community as in life:

"If you want to go fast, go alone. If you want to go far, go together"

CONTENTS

ACKNOWLEDGEMENTS

Climbing up and down the hills in Tivendens National Park in Sweden in the summer of 2023, I was pondering the journey of my PhD research that I am about to finish. Just like my hilly hike, my research has known ups and downs over the past six years. Getting one of my articles rejected and having to rewrite it completely in a different template was one of the low points. Also, not able to attend conferences in person during the COVID19 pandemic was something I hope not to relive again. However, I mainly experienced peaks during my PhD journey: the first AI Ethics & Society (AIES) in 2018 in New Orleans where I presented my first poster, attending the Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS) of the UN as a member of the Dutch delegation, being part of the World Summit AI: Dual use technology expert panel, presenting my work at the AIES 2022 conference in Oxford, the panel at Summit on Responsible Artificial Intelligence in the Military Domain (REAIM 2023) in The Hague, recording a BNR podcast and EuroISME webinar, and my yearly trips to Umeå University in Sweden to visit Virginia and her Responsible AI research group.

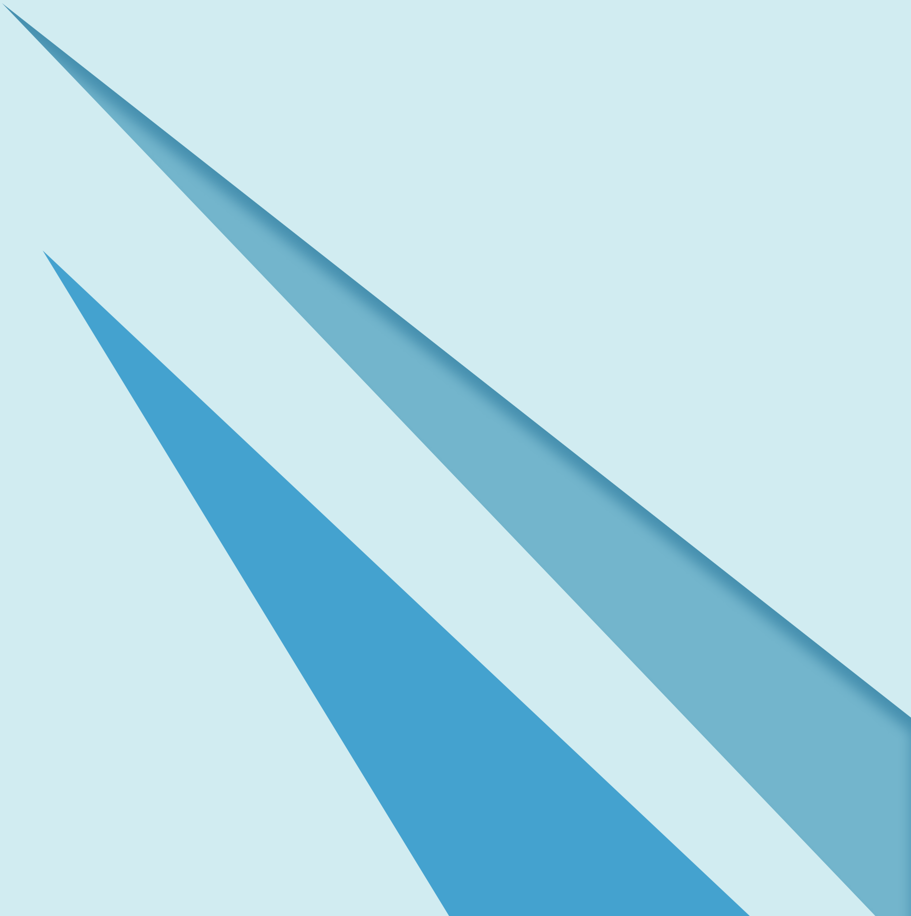
These highlights would not have been possible without the excellent supervision of my two promotors **Virginia Dignum** and **Filippo Santoni de Sio**. Thank you **Virginia**, for sharing your experience and providing guidance during my research journey, setting a high standard with your critical viewpoint and sharp comments, and for your and **Frank**'s hospitality in Umeå. You are an inspiration and role model in academia and life. **Filippo**, I greatly appreciate our discussions during the conceptual part of my research. You helped me develop my academic attitude and thinking, especially at the start of my PhD you encouraged me to find the depth in the arguments of my reasoning. Next to my two promotors, I would like to thank my fellow PhD students and researchers with whom I have worked and discussed my ideas over the past six years. **Andrea**, **Andreas**, **Klara** and **Rijk**, thank you for sharing your ideas and feedback. You made me feel part of the research community.

Lastly, I would like to thank the **Royal Netherlands Army** for supporting my PhD and allowing me to conduct independent research on this topic. Especially my **team members** and **managers** of my last two positions who had to deal with my absence and take over my work when I was doing research, visiting conferences, writing my articles and this dissertation. I hope that this dissertation on Comprehensive Human Oversight of Autonomous Weapon Systems will lead to awareness and initiates discussion on the deployment of Autonomous Weapon Systems and Responsible AI in the Defence organisation.

Part I

INTRODUCTION

In the introduction we describe the context of our research, our research objective and questions, the scenario that we use in the different phases of our research, our research approach and we conclude with the outline of our thesis.



1|

Introduction

Autonomous Weapon Systems are weapons systems equipped with Artificial Intelligence (AI). They are increasingly deployed on the battlefield (Dawes, 2023; Heather M. Roff, 2016; Tucker, 2023). Autonomous systems can have many benefits in the military domain, for example in the Ukraine where the Fortem DroneHunter F700, which is an autonomous drone with radar control and artificial intelligence, is deployed to shield the country's energy facilities from Russian attacks (Soldak, 2023). Yet the nature of Autonomous Weapon Systems might also lead to security risks and unpredictable activities as Non-Governmental Organisations (NGO's) Human Rights Watch (Human Rights Watch, 2023) and the International Committee of the Red Cross (ICRC, 2023) indicate in their statements to The Group of Governmental Experts (GGE) on emerging technologies in the area of Lethal Autonomous Weapons Systems (LAWS) of the Convention on Certain Conventional Weapons (CCW) of the United Nations. Next to security risks and unpredictable activities, the impact on human dignity and the emergence of an accountability gap are mentioned as concerns with the use of Autonomous Weapon Systems. The alleged offence to human dignity entailed in delegating life-or-death decision-making to a machine is linked to the value of human life. The Campaign to Stop Killer Robots (2023) states on their website that: '*…a machine should not be allowed to make a decision over life and death.*', because it is lacking human judgement and understanding of the context of its use. The United Nations are also voicing their concerns and state that '*Autonomous weapons systems that require no meaningful human control should be prohibited, and remotely controlled force should only ever be used with the greatest caution*' (General Assembly United Nations, 2016).

At the same time, many scholars express concerns that Autonomous Weapon Systems will lead to an "accountability gap" or "accountability vacuum"; circumstances in which no human can be held accountable for the decisions, actions and effects of Autonomous Weapon Systems (Matthias 2004; Asaro 2012; Asaro 2016; Crootof 2015; Dickinson 2018; Horowitz and Scharre 2015; Wagner 2014; Sparrow 2016; Roff 2013; Galliott 2015). This concern is also reflected in one of the guiding principles for LAWS of the GGE on emerging technologies in the area of LAWS of the CCW of the United Nations: *'Human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. This should be considered across the entire lifecycle of the weapon system.'* (UN GGE LAWS 2018).

Hence, the deployment of Autonomous Weapon Systems on the battlefield without direct human oversight is not only a military revolution according to Kaag and Kaufman (2009), but can also be considered a moral one. As large-scale deployment of AI on the battlefield seems unavoidable (Rosenberg & Markoff, 2016), the research on ethical and moral responsibility is imperative.

The concerns described above highlight that responsibility, accountability and human control are values often mentioned in the societal and academic debate on autonomous systems. Responsibility can be forward-looking to actions to come and/ or backward-looking to actions that have occurred. Accountability is a form of backward-looking responsibility that refers to the ability and willingness of actors to provide information and explanations about their actions and defines mechanisms for corporate and public governance to hold agents and organisations accountable in a forum. Responsibility contributes to minimizing unintended consequences by anticipating on actions and unintended consequences to come and taking measures to prevent or mitigate them. Accountability can decrease unintended consequences in providing information and explanations by actors of their previous actions in order for other actors to learn from them and prevent mistakes and unintended consequences of their own.

We found little empirical research that supports the concerns mentioned above or that provide insight in how responsibility and accountability regarding the deployment of Autonomous Weapon Systems are perceived by the general public and military. The Open Robots Ethics initiative surveyed the public opinion in a poll in 2015 (Open Roboethics initiative, 2015) and issued a report. However, the results were not published in an academic journal and the survey was not extensive enough to draw substantive conclusions. The notion of Meaningful Human Control is often mentioned as a requirement in the debate on Autonomous Weapon Systems to ensure accountability and responsibility over these type of weapon systems. The U.K.-based NGO Article 36 is credited for putting the concept of “Meaningful Human Control” at the centre of the discussion on Autonomous Weapon Systems by mentioning it in several reports and policy papers since 2013 (Amoroso & Tamburrini, 2021). Since then, the concept of Meaningful Human Control is often mentioned as requirement (Adams, 2001; Heather M Roff & Moyes, 2016; Vignard, 2014) to ensure accountability and responsibility for the deployment of Autonomous Weapon Systems, but this concept is not-well defined in literature and quantifying the level of control needed is hard (Schwarz, 2018). Adams (2001) noticed as early as 2001 that the role of the human changed from being an active controller to that of a supervisor and that direct human participation in decisions of AI systems would become rare. Some scholars are working on defining the concept of Meaningful Human Control in Autonomous (Weapon) Systems (Ekelhof, 2015; Horowitz & Scharre, 2015; Mecacci & Santoni De Sio, 2019; Santoni de Sio & Van den Hoven, 2018). In recent years, other scholars have been building on this work by operationalising the concept of Meaningful Human Control (see section 7.1. for emerging insights on operationalising Meaningful Human Control). Amoroso & Tamburrini (2021) bridge the gap between weapon usage and ethical principles based on ‘if-then’ rules, Umbrello (2021) proposes two Levels of Abstraction in which different agents have different levels of control over the decision-making process to deploy an Autonomous Weapon System, and Cavalcante Siebert et al. (2023), who build on the two necessary conditions for Meaningful Human Control- tracking and tracing – distinct by Santoni de Sio & Van den Hoven (2018), to create actional properties for the design of AI systems in which each of the properties human and artificial agents interact. In their reflection on their work the authors highlight that '*Meaningful human control is necessary but not sufficient for ethical AI.*' (Cavalcante Siebert et al., 2023, p. 252). The authors amplify this by stating that for a human-AI system to align with societal values and norms, Meaningful Human Control must entail a larger set design objectives which can be achieved by transdisciplinary practices.

In our opinion, Meaningful Human Control alone will not suffice as requirement to minimize unintended consequences of Autonomous Weapon Systems due to several reasons. Firstly, the concept of Meaningful Human Control is potentially controversial and confusing as human control is defined and understood differently in various literature domains (see section 2.11 for an overview of the concept of control in different domains). Secondly, standard concepts of control in engineering and the military domain entail a capacity to directly cause or prevent an outcome that is not possible to achieve with an Autonomous Weapon System, because once an autonomous weapon is launched you cannot intervene by human action. And finally, specific literature on Meaningful Human Control over Autonomous Weapon Systems does not offer a consistent usable concept. We believe that a different approach is needed to minimize unintended consequences of Autonomous Weapons Systems. Therefore, we propose an additional perspective that focusses on human oversight instead of Meaningful Human Control.

Several scholars are describing the concept of human oversight in Autonomous Weapon Systems and AI in general. HRW and IHRC (2012) state that human oversight on robotic weapons is required to guarantee adequate protection of civilians in armed conflicts and they fear that when humans only retain a limited, or no, oversight role, that they could be fading out the decision-making loop. Taddeo and Floridi (2018) describe that human oversight procedures are necessary to minimize unintended consequences and to compensate unfair impacts of AI. The European Commission mentions Human Agency and Oversight as one of the Ethics Guidelines for Trustworthy AI (European Commission, 2019). However, current human oversight mechanisms are lacking effectiveness (HRW & IHRC, 2012) and might gradually erode to become meaningless or even impossible (Williams, 2015). Marchant et al. (2011) note that several governance mechanisms can be applied to achieve human oversight of Lethal Autonomous Robots. Oversight incorporates the governance mechanisms of institutions and is therefore broader than merely Meaningful Human Control. We propose a human oversight mechanism from a governance perspective to ensure accountability and responsibility in the deployment of Autonomous Weapon Systems in order to minimize unintended consequences. In the remainder of this chapter, we will describe the research objectives, knowledge gaps and research questions that guide the development of a governance mechanism for human oversight.

1.1 RESEARCH OBJECTIVE

To ensure accountability and responsibility, a mechanism is needed to oversee and supervise the deployment of Autonomous Weapon Systems. We propose an alternative view complementary to Meaningful Human Control that incorporates the social institutional and design dimension at a governance level. This alternative view provides stakeholders additional opportunities to ensure accountability and responsibility in the deployment of Autonomous Weapon Systems. While in recent years several scholars have been working on defining the concept of Meaningful Human Control, we have found that the concept of Human Oversight is not equally studied the in literature nor a framework or implementation concept for it is offered. Also, empirical studies on the elicitation of values related to Autonomous Weapons Systems, such as accountability and responsibility, and the extent of how accountability and responsibility as values are perceived during the deployment of Autonomous Weapon Systems by common people and experts are missing. Next to this, the values of accountability and responsibility are often used interchangeably in the debate on Autonomous Weapon Systems whilst being different subjects. As stated above, responsibility contributes to minimizing unintended consequences by anticipating on actions and unintended consequences to come and taking measures to prevent or mitigate them. Accountability on the other hand can decrease unintended consequences in providing information and explanations by actors of their previous actions in order for other actors to learn from them and prevent mistakes and unintended consequences of their own. This leads to the following problem statement for this research:

> *A framework for Human Oversight is needed to ensure accountability in order to minimize unintended consequences of Autonomous Weapon Systems, but the current mechanisms for human oversight are lacking effectiveness. The concept of Meaningful Human Control will not suffice as requirement to ensure accountability in order to minimize the unintended consequences of this type of weapon system, because standard concepts of control in engineering and the military domain entail a capacity to directly cause or prevent an outcome that is not possible to achieve with an Autonomous Weapon System as once it is launched you cannot intervene by human action. Designing and implementing a framework for Human Oversight for Autonomous Weapon Systems enables proper allocation of accountability and responsibility in the deployment of Autonomous Weapon Systems.*

In taking a governance approach for ensuring accountability and responsibility in the deployment of Autonomous Weapon Systems follows a knowledge gap that is fourfold in that 1) a delineation of the values accountability and responsibility in the de debate on Autonomous Weapon Systems, 2) a theoretical account on the concept and mechanism for Human Oversight for Autonomous Weapon Systems, 3) an empirical study to elicit values and survey people's perception on accountability and responsibility during the deployment of an Autonomous Weapon System*,* and 4) a framework and implementation concept to represent criteria for Human Oversight for Autonomous Weapon Systems are lacking. These knowledge gaps can be filled by analysing the values of accountability, responsibility and the concept of Human Oversight, conducting a value elicitation study and by designing a framework and implementation concept for Human Oversight over Autonomous Weapons Systems. This leads to the following research objective:

> *To improve the allocation of accountability and responsibility by designing a framework and implementation concept such that criteria for Human Oversight are identified, represented and validated in order to minimize unintended consequences in the deployment of Autonomous Weapon Systems.*

To fulfil this research objective the following research questions need to be answered:

Q1 *What are Autonomous Weapon Systems and how are the values of accountability and responsibility related to the concerns for the deployment of Autonomous Weapon Systems?*

Q2 *How should the values of accountability, responsibility and the concept of Human Oversight be characterized?*

Q3 *Which control mechanisms are described in literature and present in the military domain, and which gaps in control mechanisms can be identified by the introduction of Autonomous Weapon Systems?*

Q4 *To what extent can an empirical study be used to elicit values and how does this lead to changes in perception of the values accountability and responsibility in a scenario of Autonomous Weapon System deployment?*

Q5 *To what extent can Human Oversight be translated into observable criteria for the deployment of Autonomous Weapon Systems?*

Q6 *To what extent can observable criteria for Human Oversight be incorporated in an implementation concept for the deployment of Autonomous Weapon Systems?*

Scientific and societal relevance

The scientific contribution of our research is twofold in that (1) our research contributes to a delineation of accountability, responsibility and Human Oversight that adds to the current body of literature, and (2) the framework and implementation concept for Human

Oversight for Autonomous Weapon Systems might also be applied to other AI fields to ensure accountability of other Autonomous Systems, such as those for Autonomous Vehicles or in the medical domain.

The societal contribution of our research is a framework and implementation concept for Human Oversight that would lead to a proper allocation of accountability in the decision-making of the deployment of an Autonomous Weapon System. By identifying the supervisor of these actions, it might be possible to attribute responsibility for the actions taken by the weapon system. This contributes to decreasing the likelihood of unintended consequences in the deployment of Autonomous Weapon Systems.

Scope

Much of the literature in the academic and societal debate on Autonomous Weapon Systems is written and discussed by legal experts and philosophers in the context of International Humanitarian Law and the Geneva Conventions which are aimed to limit the effects of armed conflicts (ICRC, 2010). As we are no legal experts, this research will stay within the boundaries and rules of the Laws of Armed Conflict (LOAC) as currently defined in the mainstream literature and we will not question these. As with any weapon system, LOAC also applies to Autonomous Weapon Systems.

Furthermore, this research will focus on the deployment of Autonomous Weapon Systems in the near future, which we define as: *within the next 15 years*. This entails that we will not study weapons equipped with Artificial General Intelligence or futuristic technology that is not possible to construct yet, but we focus on technology that is currently being developed. In this study, we will take a broad perspective on Autonomous Weapon Systems and will not limit to a specific type of weapon, like autonomous drones, but also consider types such as Autonomous Weapon Systems in the cyber domain and as part of a network of systems.

1.2 SCENARIO

In the interest of clarity and consistency, the same scenario will be used in the different phases of this research. This scenario describes a threat to soldiers which could occur during military road clearing operations to find and clear improvised explosive devices. The technology (the facial and image recognition software for people and different preprogrammed options to engage) that is described for the Autonomous Weapon Systems exists separately, but is as far as we know not yet incorporated in a deployed Autonomous Weapon System. However, due to the technological advances we expect that these technological features are possible in the near future which makes this a realistic scenario.

We have chosen to base the scenario on Airborne drones because these systems are being deployed in current conflicts, for example in Ukraine and Gaza, whilst unmanned ground-based systems are primarily in a testing phase and, as far as we know, not yet widely deployed on a battlefield. Underwater unmanned systems are also being used in current conflicts, but due to the underwater environment and lack of people in the vicinity of the system the risk on collateral damage is minimal.

The scenario reads as follows:

> *An Autonomous Weapon System provides force protection for soldiers that are clearing the road from improvised explosive devices. The Autonomous Weapon System is equipped with surveillance equipment, weapons (air-to-ground missiles) and flies autonomously in the Area of Operation. It is programmed to avoid flying over a restricted operating zone and an electronic warfare threat. The Autonomous Weapon System is equipped with facial and image recognition software for people, weapons and explosives. It is programmed with different options to engage when it recognizes a threat to the soldiers that are clearing the road. The Autonomous Weapon System detects movement behind a large rock near a narrow part of the road at a distance of 300 meters of the road clearance soldiers.*

1.3 RESEARCH APPROACH

In this research, we apply the Value-Sensitive Design (VSD) method as research approach. The VSD is a three-partite approach that allows for considering human values throughout the design process of technology (Figure 1). It is an iterative process for the conceptual, empirical and technological investigation of human values implicated by the design (Davis & Nathan, 2015; Friedman & Kahn Jr, 2003). The conceptual investigation consists of two parts: (1) identifying the direct stakeholders, i.e. those who will use the technology, and the indirect stakeholders, i.e. those whose lives are influenced by the technology, and (2) identifying and defining the values that the use of the technology implicates. The empirical investigation looks into the understanding and experience of the stakeholders in a context relating to the technology and implicated values will be examined. In the technical investigation, the specific features of the technology are analysed (Davis & Nathan, 2015). The VSD can be used as a roadmap for engineers and students to incorporate ethical considerations into the design (Cummings, 2006).

There has been some critique voiced regarding the VSD approach. One of the concerns Davis and Nathan (2015) mention is that the VSD posits that certain values are universal, but that these may differ based on culture and context. A response to counter this would be to take an empirical basis for one's viewpoint instead of a philosophical one, or acknowledge that the researcher's position is not the only valid position to be considered (Borning & Muller, 2012). Borning and Muller (2012) pose a pluralistic position in that the VSD should not recommend either a universal or a relative view on values, but it should leave engineers free to decide which view is most appropriate in context of their design. Also, while moral values can and do differ across cultures, some values guiding basic principles of international law – e.g. human rights and protection of civilians- have been formally endorsed by countries with different histories and cultures (ICRC, 2010).

In line with Borning and Muller (2012) we used the VSD approach in our research as guidance and not as a goal in itself. In the conceptual phase, we slightly deviate from the original VSD method, because we do not conduct a full stakeholder analysis to identify the stakeholders in the conceptual investigation phase, but we focus on the obvious stakeholder groups; military, policymakers, industry and Non-Governmental Organisations (NGO's). For the identification of values in step 2, we used our previous work (Verdiesen, Santoni de Sio, & Dignum, 2019) in which we researched the values related to Autonomous Weapon Systems. In the technical investigation phase, we do not design an Autonomous Weapon System as one intuitively might expect, because this would be an immense project well beyond the scope of this research. Yet, we used a discrete-event modelling language (Coloured Petrinets (CPNs)) for modelling synchronisation concurrency and communication processes. We created a model that represents observable criteria of a pre-flight mission planning and post-flight mission evaluation process for autonomous drones.

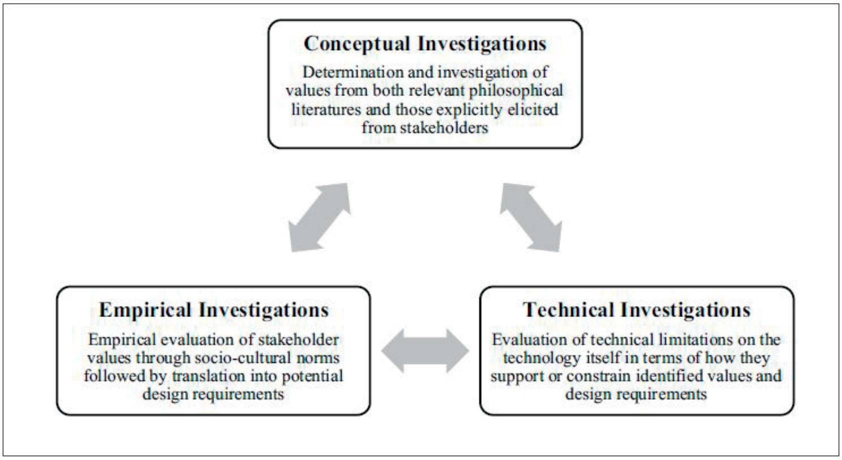

Conceptual Investigations
Determination and investigation of values from both relevant philosophical literatures and those explicitly elicited from stakeholders
Empirical Investigations
Empirical evaluation of stakeholder values through socio-cultural norms followed by translation into potential design requirements
Technical Investigations
Evaluation of technical limitations on the technology itself in terms of how they support or constrain identified values and design requirements

Figure 1: VSD (as in: Umbrello & Van de Poel, 2021)

1.4 OUTLINE OF THESIS

This thesis consists of five parts of which the introduction is part I. The remainder of this thesis is structured according to the phases of the Value-Sensitive Design approach and reads as follows:

Part II: Conceptual investigation phase
In chapter 2 the relevant literature on decision-making processes in AI, architectures for ethical decision-making in AI, autonomy, Autonomous Weapon Systems, values, values related to Autonomous Weapon Systems and value hierarchy as a Design for Values approach is reviewed. Parts of this chapter have been published in Verdiesen (2017) and Verdiesen, De Sio, and Dignum (2019).

In chapter 3 we present the Comprehensive Human Oversight Framework by describing the layers and the connections between them and identifying gaps in the control mechanisms. To mitigate these gaps, we applied the Glass Box framework on the Comprehensive Human Oversight Framework. We conclude chapter 3 by closing the gap from the review stage back to the interpretation stage by means of a feedback process. Parts of this chapter have been published in Verdiesen, De Sio, and Dignum (2019) and Verdiesen, Aler Tubella, and Dignum (2021).

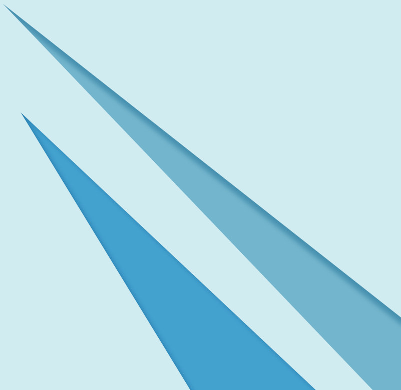
Part III: Empirical investigation phase
In chapter 4 we describe the empirical investigation phase of our research which consists of conducting expert interviews, the Value Deliberation Process as a means to elicitate values and validating the results by consulting experts. For reflection and validation, we discussed the Comphrensive Human Oversight Framework and aspects of drone deployments during interviews and an extra round of validation was conducted by inviting experts- who had not been part of the expert panel- to reflect on the findings of the value elicitation. Parts of this chapter have been published in Verdiesen and Dignum (2022).

Part IV: Technical investigation phase
In chapter 5 we present the implementation concept for operationalising the Glass Box framework. After introducing the scenario, we describe Coloured Petri Nets: a discrete-event language for modelling synchronisation concurrency and communication processes that we used to model the implementation concept. We conclude with remarks on validating the implementation concept. Parts of this chapter have been published in Verdiesen, Aler Tubella, and Dignum (2021).

Part V: Conclusion and discussion

In chapter 6 we follow the three phases of our research approach to answer our research questions based on the results of our research to conclude if our research objective - to improve the allocation of accountability and responsibility in the deployment of Autonomous Weapon Systems by designing a framework and implementation concept such that the criteria for Human Oversight are identified, represented and validated- is reached.

Chapter 7 contains the discussion on our research in which we highlight the emerged insights over the past five years on the definition of Autonomous Weapon Systems and the operationalisation of Meaningful Human Control, followed by the limitations of this research and suggestions for future work. We conclude this chapter by presenting the contributions and recommendations of our research.



Part II

CONCEPTUAL INVESTIGATION PHASE

This part on the conceptual investigation phase of our research consists of two chapters; 1) we review relevant literature on decision-making processes in AI, architectures for ethical decision-making in AI, autonomy, Autonomous Weapon Systems, values, values related to Autonomous Weapon Systems and value hierarchy in the Design for Values approach and 2) we present the Comprehensive Human Oversight Framework by describing the layers and the connections between them, identifying gaps in the control mechanisms and describe a feedback process to close the gaps. Parts of chapter 2 has been published in Verdiesen (2017) and Verdiesen et al. (2019).

2|

Extensive literature review

In this chapter we review relevant literature on decision-making processes in AI, architectures for ethical decision-making in AI, autonomy, Autonomous Weapon Systems, values, values related to Autonomous Weapon Systems, a value hierarchy as a Design for Values approach, responsibility, accountability and accountability gaps, perspectives on control and human oversight.

2.1 DECISION-MAKING PROCESSES IN AI

Decision-making processes in Artificial Intelligence (AI) have been studied for over two decades and is quite well delineated in AI and engineering literature (see Table 1 in appendix C for an overview). Decision-making is defined as a process in which: '*an entity is in a situation, receives information about that situation, and selects and then implements a course of action.*' (Miller, Wolf, & Grodzinsky, 2017, p. 390). Adams (2001) noticed as early as 2001 that the role of the human changed from being an active controller to that of a supervisor, and that direct human participation in decisions of AI systems would become rare. The concept of adjustable autonomy, i.e., switching between autonomy levels, is mentioned often in literature to deal with changes in context, the need of the operator and the control humans can exert over the machine (Cordeschi, 2013; Côté, Bouzid, & Mouaddib, 2011; van der Vecht, 2009).

As is noted by Cordeschi (2013), optimal choices in decision-making for humans and AI do not exist, therefore only satisficing choices can be made. It depends on the situation if humans or AI can make the most reliable decision. In order for an AI system to be able to make ethical decisions it is not necessary that its decision-making is similar to that of a human, but the system will need a mechanism such as a heuristic algorithm to analyse its past decisions and prepare for future decisions (Miller et al., 2017). However, moving from a technical debate to an ethical point of view, according to Kramer, Borg, Conitzer, and Sinnott-Armstrong (2017), the question is not only if we can build moral decision-making in AI, but also if 'moral AI' systems should be permitted at all to make decisions. While this is certainly an important question, it is interesting to note that, as a matter of fact, people's moral intuitions about this issue appears to be highly dependent on their acquaintance with computers. It seems that the more people are familiar with computers, the more they prefer decisions made by computers over decisions made by humans (Araujo, Helberger, Kruikemeier, & De Vreese, 2020; Kramer, Borg, Conitzer, & Sinnott-Armstrong, 2017). Araujo et al. (2020) found that for high impact decisions, the potential fairness, usefulness and risk of specific decision-making automatically by AI compared to human experts was often on par or even better evaluated. Based on their research, Kramer et al. (2017) expect that the more people gain experience with computer decision-making and it becomes more visible, the more it will be accepted by the general public.

2.2 ARCHITECTURES FOR ETHICAL DECISION-MAKING IN AI

When computer programs of autonomous systems are implemented in the unpredictable real-world, the behaviour of these systems becomes non-deterministic and a range of possible outcomes can occur (Dennis, Fisher, Slavkovik, & Webster, 2016). To govern these unpredictable outcomes of autonomous systems in real-world scenarios, a mechanism is needed to influence the agent's (ethical) decision-making. In engineering literature, two types of architectures for ethical decision-making of AI can be found (see Table 2 in appendix C for an overview).

The first is based on an 'ethical layer' that governs the behaviour of the agent from outside the system. Arkin, Ulam, and Wagner (2012) designed and implemented an 'ethical governor' that consists of 2 processes; 1) ethical reasoning that transforms incoming perceptual, motor and situational awareness data into evidence, and 2) constraint application that uses the evidence to apply constraints based on Laws Of War and Rules Of Engagement to suppress unethical behaviour when applying lethal force. Dennis et al. (2016) proposes a hybrid architecture in which reasoning is done by a rational BDI [Beliefs, Desires and Intentions] agent. Based on this framework the agent selects plans from a given ethical policy which is the most ethical plan available based on its beliefs. Earlier work by Li et al. (2002) consists of a hierarchical control scheme developed to enable multiple Unmanned Combat Air Vehicles (UCAVs) to autonomously achieve demanding missions in hostile environments. The scheme consists of four layers: 1) a high-level path planner, 2) a low-level path planner, 3) a trajectory generator and 4) a formation control algorithm. More recently, Vanderelst and Winfield (2018) designed an additional or substitute framework for implementing robotic ethics as alternative for logic-based AI that currently dominates the field. They implemented ethical behaviour in robots by simulation theory of cognition in which internal simulations for actions and prediction of consequences are used to make ethical decisions. The method is a form of robot imagery and does not make use of verification of logical statements that is often used to check if actions are in accordance with ethical principles.

The second type of architecture for ethical decision-making of AI is logic based. This type derives logical rules from natural language and applies the rules to the system to govern its ethical behaviour. Anderson, Anderson, and Berenz (2016) describe a *case-supported principle-based behavior paradigm* (CPB) to govern an elderly care robot's behaviour. The system uses principles, that are abstracted from cases, that have consensus of ethicists, to choose its next action. It sorts the actions by weighing them according to ethical preferences, which are based on values, and selects the action that is highest ranked. Another formal approach is HERA (Hybrid Ethical Reasoning Agents) which is a software library to model autonomous moral decision-making (Lindner, Bentzen, & Nebel, 2017). HERA represents the robot's possible actions together with the causal chains of consequences the actions initiate. Logical formulae are used to model ethical principles. The software library implements several ethical principles or interpretation of ethical principles, such as the principle of Double Effect, utilitarianism and a Pareto-inspired principle. The applied format is called a causal agency model. It reduces determining moral permissibility by checking if principle-specific logical formulae are satisfied in a causal agency model. Recent work of Bonnemains, Saurel, and Tessier (2018) demonstrates a formal approach is developed to link ethics and automated reasoning in autonomous systems. The formal tool models ethical principles to compute a judgement of possible decisions in a certain situation and explains why this decision is ethically acceptable or not. The formal model can be used on utilitarian and deontological ethics and the Doctrine of Double effect to examine the results generated by these three different ethical frameworks. They found that the main challenge lies in formalizing philosophical definitions in natural language and to translate them in generic computer programmable concepts that can be easily understood and that allows for ethical decisions to be explained.

2.3 AUTONOMY

The notion of autonomy is a not well-defined and often misunderstood concept. Nowadays in the context of AI, autonomy is often a synonym for Machine Learning, an example can be found in Melancon (2020), but autonomy encompasses much more than that. Castelfranchi and Falcone (2003) define autonomy as a notion that involves relationships between three entities: a) the main subject *x*, b) the goal *μ* that must be obtained by the main subject *x* and c) a second subject *y* upon the main subject *x* is autonomous. This is expressed in the statement: "*x* is autonomous about *μ* with respect to *y*". For example, if *x* is an autonomous drone, its autonomy implies that the autonomous drone *x* can autonomously decide on the travel route (the goal *μ*) given a destination (i.e. GPS coordinates) set by its operator *y*. Three type of autonomy relationships can be identified based on this description: (1) *executive autonomy; x* is autonomous in its means instead of it goals, which is the case of the example of the autonomous drone, (2) *goal autonomy*; *x* can set its goals on its own, and (3) *social autonomy*; *x* can execute its goals by itself without other agents (Castelfranchi & Falcone, 2003).

Wooldridge and Jennings (1995, p. 116) also refer to autonomy in their list of four properties for defining an agent: '*1) autonomy: agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state (Castelfranchi, 1995), 2) social ability: agents interact with other agents (and possibly humans) via some kind of agent-communication language (Genesereth*

& Ketchpel, 1994), 3) reactivity: agents perceive their environment (which may be the physical world, a user via a graphical user interface, a collection of other agents, the Internet, or perhaps all of these combined), and respond in a timely fashion to changes that occur in it; and 4) pro-activeness: agents do not simply act in response to their environment, they are able to exhibit goal-directed behaviour by taking the initiative.'

In their article on defining Autonomous Weapon Systems, Taddeo and Blanchard (2022) delineate and specify the difference between automatic/automated and autonomous agents. They state: '*The ability of an artificial agent to change its internal states without the direct intervention of another agent marks (binarily) the line between automatic/automated and autonomous. A rule-based artificial system and a learning one both qualify as autonomous following this criterion.*'(Taddeo & Blanchard, 2022, p. 17). An automated system on the other hand can perform a complex and a predetermined task. A robot in a car manufacturing factory is an example of an automated system. The authors also state that it is increasingly more common that adaptability is a key characteristic for Autonomous Weapon Systems, which will be their potential to deal with complex and fast pacing scenarios, but also will also cause unpredictability, lack of control and transparency, and responsibility gaps (Taddeo & Blanchard, 2022). Taddeo & Blanchard (2022) base their delineation on the work of Floridi and Sanders (2004) who describe three criteria for intelligent systems:

> *'(a) Interactivity means that the agent and its environment (can) act upon each other. Typical examples include input or output of a value, or simultaneous engagement of an action by both agent and patient – for example gravitational force between bodies.*
>
> *(b) Autonomy means that the agent is able to change state without direct response to interaction: it can perform internal transitions to change its state. So an agent must have at least two states. This property imbues an agent with a certain degree of complexity and independence from its environment.*
>
> *(c) Adaptability means that the agent's interactions (can) change the transition rules by which it changes state. This property ensures that an agent might be viewed, at the given LoA [Level of Abstraction], as learning its own mode of operation in a way which depends critically on its experience. Note that if an agent's transition rules are stored as part of its internal state, discernible at this LoA, then adaptability follows from the other two conditions.'* (Floridi & Sanders, 2004, pp. 357-358).

Above, autonomy is described from an engineering perspective, but it can also be viewed from a human value perspective. For instance, in Bioethics, which describes the values that are important as guiding principles in the medical field, autonomy is defined as *acting intentionally without controlling influences that would mitigate against a voluntary act* (Beauchamp and Walters, 1999). The definition of autonomy in the field of AI should be kept distinct from the definition of human autonomy and its moral value, because they do not represent the same constructs. Although autonomy is an important human value which will be useful in the next section, it is less relevant from an engineering perspective to interpret autonomy as a singular construct for a technical system, because weapon systems may comprise of different levels of autonomy. But even in the case of a "fully Autonomous Weapon System", *'[…] that, without human intervention, selects and engages targets matching certain predefined criteria, following a human decision to deploy the weapon on the understanding that an attack, once launched, cannot be stopped by human intervention*.' (AIV & CAVV, 2015; Broeks et al., 2021) the type of autonomy can at most be executive autonomy, because a human will set its goals and the weapon will not decide on its goals or deployment itself. Also, the context will constrain the autonomy of a "fully Autonomous Weapon System" as autonomous systems are created with task goals and boundary conditions (Bradshaw, Hoffman, Woods, & Johnson, 2013). In case of Autonomous Weapon Systems, the context might include physical limitations to the area of operations, for example the presence, or lack of, civilians in the land, sea, cyber, air or space domain. In the next section, several definitions of Autonomous Weapons Systems will be provided and the rationale for choosing the definition of the AIV and CAVV (2015) mentioned above is given.

2.4 AUTONOMOUS WEAPON SYSTEMS

Although the societal and academic debate on Autonomous Weapon Systems has drawn a lot of attention in the recent years, we found that the topic was not well delineated in the academic literature. We start this subsection with an overview of the many different definitions and present two classifications of Autonomous Weapon Systems to conclude this section.

Definition

Autonomous Weapon Systems are an emerging technology and there is still no internationally agreed upon definition (AIV & CAVV, 2015; Sayler, 2021). Even consensus if Autonomous Weapon Systems should be defined at all is lacking. Although some scholars provide definitions in their writings (see Table 3 in appendix C), others caution against such a specification. NATO states that: '*Attempting to create definitions for "autonomous systems" should be avoided, because by definition, machines cannot be autonomous in* *a literal sense. Machines are only "autonomous" with respect to certain functions such as navigation, sensor optimization, or fuel management.'* (Kuptel & Williams, 2014, p. 10). The United Nations Institute for Disarmament Research (UNDIR) is also cautious about providing a definition of Autonomous Weapon Systems, because they argue that the level of autonomy depends on the '*critical functions of concern and the interactions of different variables*' (UNDIR, 2014, p. 5). They state that one of the reasons for the differentiation of terms regarding Autonomous Weapon Systems is that sometimes things (drones or robots) are defined, but in other times a characteristic (autonomy), variables of concern (lethality or degree of human control) or usage (targeting or defensive measures) are drawn into the discussion and become part of the definition. In a recent paper, Taddeo and Blanchard (2022) describe twelve definitions of (Lethal) Autonomous Weapon Systems provided by States and international organisations. They provide a value neutral definition of Autonomous Weapon Systems of their own (see Table 3 in appendix C).

Various definitions of Autonomous Weapon Systems are listed in Table 3 in appendix C. Some authors use the term military robots which have a certain level of autonomy. As military robots can be viewed as a subclass of Autonomous Weapon systems according to the classification of Royakkers and Orbons (2015) (Figure 2) we included them in the list of definitions. In our opinion the definition in the report of the ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS (AIV & CAVV) captures the description of Autonomous Weapon Systems best from an engineering and military standpoint, because it takes predefined criteria into account and is linked to the military targeting process as the weapon will only be deployed after a human decision. In their 2021 report on Autonomous Weapons Systems the AIV & CAVV continue to use this definition (Broeks et al., 2021). Therefore, we will follow this definition and define Autonomous Weapon Systems as:

> *'A weapon that, without human intervention, selects and engages targets matching certain predefined criteria, following a human decision to deploy the weapon on the understanding that an attack, once launched, cannot be stopped by human intervention.'*(AIV & CAVV, 2015, p. 11; Broeks et al., 2021, p. 11).

Classification of Autonomous Weapon Systems

Not only are Autonomous Weapon Systems ambiguously defined, they also have not been uniformly classified. We present two classifications in this subsection. Royakkers and Orbons (2015) describe several types of Autonomous Weapon Systems (Figure 2) distinct between (1) *Non-Lethal Weapons* which are weapons *'…without causing (innocent) casualties or serious and permanent harm to people.'* (Royakkers & Orbons, 2015, p. 617), such as the Active Denial System which uses a beam of electromagnetic energy to keep people at a certain distance from an object or troops, and (2) *Military Robots* which they define *'…as reusable unmanned systems for military purposes with any level of autonomy.'* (Royakkers & Orbons, 2015, p. 625). Military robots are subdivided in three categories; vehicles that are ground based, for example for unmanned reconnaissance and clearing road bombs, vehicles that can navigate unmanned on or below the water surface, such as a gun-station on a ship or an autonomous submarine, and vehicles that are unmanned combat aerial vehicles (UCAV's). These UCAV's are classified by Royakkers and Orbons (2015) as tele-operated, of which 'drones' are the most well-known example, and autonomous UCAV's, which are gradually developed by the US Department of Defense (Rosenberg & Markoff, 2016).

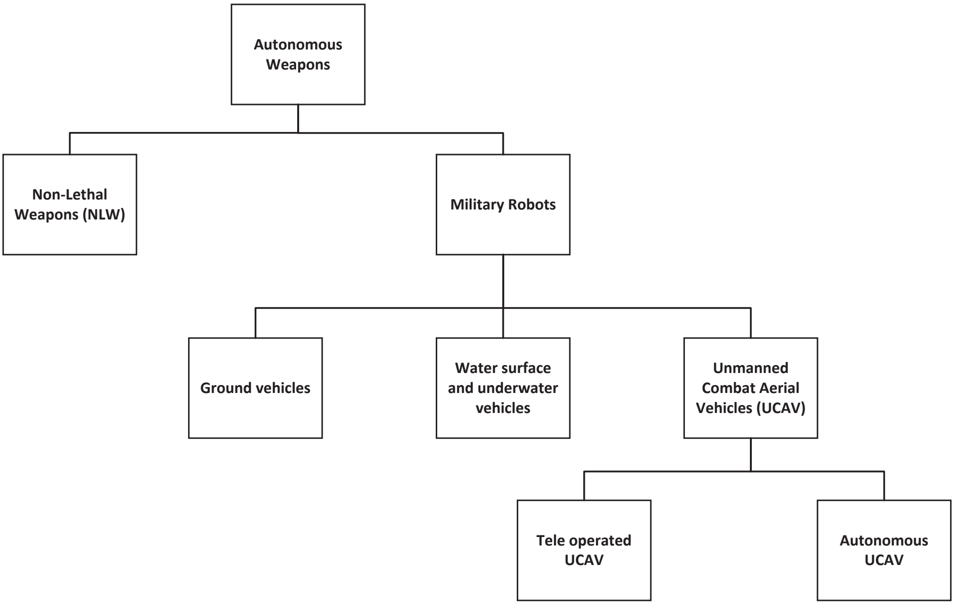


Figure 2: Classification of Autonomous Weapon Systems based on Royakkers and Orbons (2015)

Galliott (2015) provides another type of classification of Autonomous Weapon Systems based on four levels of autonomy for unmanned systems:

1. Autonomy level 1 – Non-autonomous/ teleoperated: *'A human operator controls each and every powered movement of the unmanned platform. Without the operator, teleoperated systems are incapable of effective operation.'*
2. Autonomy level 2 – Supervisory Autonomy: *'A human operator specifies movements, positions or basic actions and the system then goes about performing these. The operator must provide the system with frequent input and diligent supervision in order to ensure correct operation.'*

3. Autonomy level 3 – Task Autonomy: ‘*A human operator specifies a general task and the platform processes a course of action and carries it out under its own supervision. The operator typically has the means to oversee the system, but this is not necessary for the operation*.’
4. Autonomy level 4 – Full Autonomy: ‘*A system with full autonomy would create and complete its own tasks without the need for any human input, with the exception of the decision to build such a system. The human is so far removed from the loop that the level of direct influence is negligible. These systems might display capacities that imitate or replicate the moral capacities of sentient human beings (though no stand on this matter shall be taken here)*’ (Galliott, 2015, p. 7).

This classification is in our opinion a good attempt in classifying the degree of autonomy of Autonomous Weapon Systems, but we have some reservations from an engineering point of view. Galliott (2015) himself states that it would be possible to merge the second and third level of autonomy, because both are a semi-autonomous operational level. We agree with his statement, but this is not the main issue we have with these definitions. We believe that it is odd to start list of autonomy levels with a category of non-autonomous systems. More importantly, in the fourth level of autonomy the author states that: *‘these systems might display capacities that imitate or replicate the moral capacities of sentient human beings’*. It seems he refers to the definition of strong or general AI, in that a computer has cognitive states and programs can explain human cognition (Searle, 1980). To state that an autonomous system possesses moral capacities shows in our opinion a lack of technical knowledge on current AI systems as these are not more than computers that display Interactivity, Autonomy and Adaptability features (Floridi & Sanders, 2004).

As it remains to be seen if AI capable of ‘*moral capacities of sentient human beings’* (Galliott, 2015, p. 7) will ever be developed, we believe that the classification Galliott (2015) provides is not realistic with the current state of technology. The classification of Royakkers and Orbons (2015) is based on a combination of the system’s usage (e.g. ground, underwater, air) and in lesser degree the level of supervision (e.g. teleoperated or autonomous) and it displays good insight in the current and (near) future military technology. The classification of Galliott (2015) describes the degree of human supervision of the weapon system and by this takes a human-centric approach. A human-centric approach provides a good starting point to study the broader concept of Human Oversight. For this, we will explore human values and value theories to get a grasp of what people find important in life.

2.5 VALUES

Contrary to the topic of Autonomous Weapons, the concept of values has been studied extensively in the fields of Moral Philosophy and Psychology. This section presents a definition of values as used in this research, followed by an overview of theories that describe universal values, an overview of the values related to Autonomous Weapons and concludes with a value hierarchy.

Value theories

Value Theories are well-studied in the fields of Moral Philosophy and (Moral) Psychology. Moral Philosophy has a long and rich history in examining values and in this field theoretical questions are asked to investigate the nature of value and goodness (Schroeder, 2016). Often a distinction is made between instrumental values, which means there is reason to favour it for its effect that can lead to good things (Rønnow-Rasmussen, 2002), and intrinsic values, which *'…is a kind of value such that when it is possessed by something, it is possessed by it solely in virtue of its intrinsic properties*.' (Bradley, 2006, p. 112). Although Moral Philosophy is mainly concerned with theories of what 'ought to be' and is in a strict sense unaffected by empirical results (Alfano & Loeb, 2014), one of its branches: Applied Ethics is relevant for our study, because Applied Ethics bridges the abstract ethical theories and moral practice. In this study, we choose not to use the theoretical Value Theories of Moral Philosophy, but turn to the fields of Moral Psychology and Applied Ethics to get an empirical view in order to get insight into the 'is' situation instead of what 'ought to be'.

Literature in Moral Psychology differentiates values from attitudes, needs, norms and behaviour in that they are a belief, lead to behaviour that guides people and are ordered in a hierarchy that shows the importance of the value over other values (Schwartz, 1994). Values are used by people to justify their behaviours and define which type of behaviours are socially acceptable (Schwartz, 2012). They are distinct from facts in that values do not only describe an empirical statement of the external world, but also adhere to the interests of humans in a cultural context (Friedman, Kahn Jr, Borning, & Huldtgren, 2013). Values can be used to motivate and explain individual decision-making and for investigation of human and social dynamics (Cheng & Fleischmann, 2010).

Many definitions of values exist. For example, Schwartz (1994, p. 21) describes values as: '*desirable transsituational goals, varying in importance, that serve as guiding principles in the life of a person or other social entity.*'. This is quite a specific description compared to Friedman et al. (2013, p. 57) who define values as: '…*what a person or group of people consider important in life*.'. The existing definitions have been summarized by Cheng and Fleischmann (2010, p. 2) in their meta-inventory of values and they state that: …'*values* *serve as guiding principles of what people consider important in life'*. Although a rather simple description, we think it captures the description of a value best, because it combines several definitions in one using the main characteristics of values. Therefore, we will adhere to the definition of Cheng and Fleischmann (2010) in our study.

Universal values

Some research suggests that people across cultures identify with basic values which can be considered as universal human values (Friedman, Kahn Jr, Borning, & Huldtgren, 2013; Graham et al., 2012; Schwartz, 2012). Although individuals differ in attribution of importance of the values, there seems to be a surprisingly high consensus across cultures on the hierarchical order of the values (Schwartz, 2012). As part of their research some researchers created so called value inventories, which are lists of items that can be used to categorise the analysis of human values and are often accompanied by a descriptive tool for discussions on these values (Cheng & Fleischmann, 2010). The most common and well-studied value inventories are those of Schwartz (1994), Friedman et al. (2013), Beauchamp and Walters (1999) and Graham et al. (2012). The number of universal values found by researchers varies greatly. An overview of these value inventories is displayed in Table 4 in appendix C and the theories will be briefly described in the next paragraph.

Based on extensive empirical research, Schwartz (1994) mentions 10 distinct motivational types of values that are subdivided in a more fine-grained list of 56 value items which he uses to survey the 10 overarching universal values. In their description of the Value-Sensitive Design approach, Friedman et al. (2013) mention 12 values of which the first 9 are based on consequentialists and deontological moral orientations and the last 3 are chosen from the field of Human Computer Interaction (HCI) field. Graham et al. (2012) uses the term 'foundation' to describe the 5 distinct values that specify the universality of human moral nature that Haidt and Joseph (2004) use as basis of the Moral Foundation Theory. Gouveia, Milfont, and Guerra (2014) drafted a framework based on many value theories, such as Schwartz (1994) and Maslow (1943) hierarchy of needs. In the framework, the authors place the value on two dimensions; (1) with actions that drive human behaviour which can be personal, central or social goals, and (2) motivators that represent human needs which can split into thriving and survival needs (Gouveia et al., 2014).

Values are not only described in theory from a psychological perspective as outlined in the previous paragraph, but have also been practically implemented and used by means of Applied Ethics to professional domains. For example, in the medical field, which uses BioEthics to describe the values that are important as guiding principles for biomedical professionals, such as physicians, nurses and health workers. Beauchamp and Walters (1999) describe 4 values as basis for the framework of BioEthics: 1) *Autonomy*: acting intentionally without controlling influences that would mitigate against a voluntary act, 2) *Beneficence*: providing benefits for society as a whole, 3) *Justice*: being fair and reasonable and 4) *Non-maleficence*: not intentionally imposing risk or harm upon another.

Based on our literature review, we selected two value theories for our previous study (Verdiesen, 2017); one derived from the Psychological literature and the other based on Applied Ethics which is a practical application of Moral Philosophy. The first theory we selected is that of Cheng and Fleischmann (2010), because in their meta-inventory of human values they created a comprehensive list of 16 human values that is based on the values found in 12 separate studies. In our opinion, this meta-analysis captures the most important values listed by other researchers and it is an empirical example derived from the psychological literature. The second Value Theory we selected is an example of Applied Ethics that has been extensively practiced in the medical domain for over forty years. We investigated its applicability to Autonomous Weapon Systems, because the BioEthics principles address many concerns that people might have regarding Autonomous Weapon Systems.

2.6 VALUES RELATED TO AUTONOMOUS WEAPON SYSTEMS

Values as described in the value theories in section 2.5 are not often explicitly mentioned in the literature on Autonomous Weapon Systems, but the studies mentioned in Table 5 in appendix C discuss different values or related ethical issues related to Autonomous Weapon Systems. Two public reports of Human Rights Watch mention the lack of human emotion, accountability, responsibility, lack of human dignity and harm as values related to Autonomous Weapon Systems (Docherty, 2012, 2015). Sharkey and Suchman (2013) state that the values of accountability and responsibility are important to consider in the design of Robotic Systems for military operations. De Ágreda (2020) studied the CCW's proposal of Guiding Principles on Lethal Autonomous Weapon Systems and the values beneficence/ relative beneficence, human dignity, fairness and Meaningful Human Control are mentioned in these principles.

In the field of Military Ethics, Johnson and Axinn (2013) list responsibility, reduction of human harm, human dignity, honour and human sacrifice as values in their discussion on if the decision to take a human life should be handed over to a machine or not. Cummings (2006b) in her case study of the Tactical Tomahawk missile, looks at the universal values proposed by Friedman and Kahn Jr (2003) and states that next to accountability and informed consent, the value of human welfare is fundamental core value for engineers when developing weapons as it relates to the health, safety and welfare of the public.

She also mentions that the legal principles of proportionality and discrimination are important to consider in the context of weapon design. Proportionality refers to the fact that an attack is only justified when the damage is not considered to be excessive. Discrimination means that a distinction between combatants and non-combatants is possible (Hurka, 2005). Asaro (2012) also refers to the principles of proportionality and discrimination and states that Autonomous Weapon Systems open-up a moral space in which new norms are needed. Although he does not explicitly mention values in his argument, he does refer to the value of human life and the need for humans to be involved in the decision of taking a human life. Other studies primarily describe ethical issues, such as preventing harm, upholding human dignity, security, the value of human life and accountability (Horowitz, 2016; UNDIR, 2015; Walsh & Schulzke, 2015; A. P. Williams, Scharre, & Mayer, 2015).

In a previous study we identified the values that people associate with Autonomous Weapon Systems (Verdiesen, 2017; Verdiesen et al., 2019). The overview is derived from both validated value theories as from experts who are involved in the debate on Autonomous Weapon Systems or work in the military domain. After conducting two pilot studies, we selected the values *blame*, *trust, harm, human dignity, confidence, expectations, support, fairness* and *anxiety* to be incorporated in the final questionnaire of the study. The results provide insight in how military personnel and civilians working at the Dutch Ministry of Defense (MOD) perceive these values for both a Human Operated drone, as an example of current technology, as for Autonomous Weapon Systems, as future technology. To our knowledge this study is the first to empirically investigate these values related to Autonomous Weapon Systems and to compare how these values are perceived in current and future weapon systems.

Our results show that military personnel and civilians working at the Dutch MOD are more anxious about the deployment Autonomous Weapon Systems than the deployment of Human Operated drones. They also perceive them to have less respect for the dignity of human life than Human Operated drones. *Human dignity* and *anxiety* are two values that are mentioned often by the experts in their interviews so it would be essential to address these when debating the ethics of the deployment of Autonomous Weapon Systems. Our findings show that the *trust, confidence* and *support* for Autonomous Weapon Systems is lower than for Human Operated drones. We would like to note at this point that Autonomous Weapon Systems not only have drawbacks, but also have clear military advantages (Etzioni & Etzioni, 2017) and designing features to increase the trust and confidence of Autonomous Weapon Systems is beneficial from a military point of view.

2.7 DESIGN FOR VALUES

Design for Values is an approach to develop technology based on moral and societal values (van den Hoven, Vermaas, & van de Poel, 2015). It is aimed at countering the standard practice of designing technology as an alleged value-neutral artifact that, as a matter of fact, meets the requirements set by producers, clients or users and by this disregarding values of society at large. Design for Values attempts to prevent the Collingridge dilemma (Collingridge, 1980). The Collingridge dilemma implies that in early stages of technology development there is much possibility to change the design, but the information about the unintended or undesired outcomes of using the technology is scarce, while in later stages of technology development this information is available but changing the design is often impossible or expensive. In addition to be morally and socially desirable, Design for Values can have economic benefits as it contributes to the acceptability and success of innovations.

Several Design for Value approaches exist, for example the Value-Sensitive Design method (described in section 1.3), Technology Assessment, Constructive Technology Assessment and a value hierarchy. All these methods have three criteria in common: 1) the belief that values can be incorporated into technology, 2) it is morally significant to think about values in technology explicitly, and 3) in order to make a difference, value considerations need to be incorporated early on into the design process (van den Hoven et al., 2015). As an example of a Design for Values approach, we describe the theory of the value hierarchy method and apply it to the case of Autonomous Weapon Systems in the next section.

Value hierarchy

One approach to consider which values are relevant in the design of Autonomous Weapon Systems is the translation of values into design requirements which can be made visible by means of a value hierarchy (Van de Poel, 2013). This hierarchical structure of values, norms and design requirements makes the value judgements, that are required for the translation, explicit, transparent and debatable. To do so, the values that are described in the natural language will need to be translated to 'formal values in a formal language' (Aldewereld, Dignum, & Tan, 2015, p. 835). One way of formalizing values into norms would be to use a convention of rules which are represented as: ' ”*X counts as Y” or “X counts as Y in context C”* ' (Searle, 1995, p. 28). The explicitly of values in formal rules allows for critical reflection in debates and pinpoint the value judgements that are disagreed on. Transparency is important as Van de Poel (2013, p. 265) eloquently states: '*Although transparent choices are not necessarily better or more acceptable, transparency seems a minimal condition in a democratic society that tries to protect or enhance the moral autonomy of its citizens, especially in cases that design* *impacts the lives of others besides the designers, as is often the case*'.

The top level of a value hierarchy consists of the *values*, as depicted in Figure 3, the middle level contains the *norms*, which can be capabilities, properties or attributes of the artefact, and the lower level are the design *requirements* that can be identified. The relation between the levels is not deductive and can be constructed top-down, by means of specification, or bottom-up by seeking for the motivation and justification of the lower-level requirements. The bottom-up conceptualisation of values is a philosophical activity which does not require specific domain knowledge and the top-down specification of values requires context or domain specific knowledge that adds content to the design (Van de Poel, 2013).

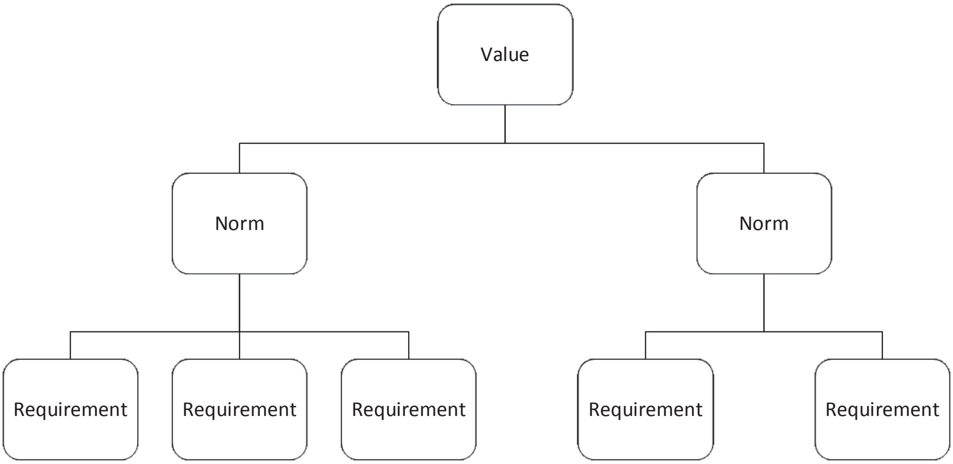


Figure 3: Conceptual model value hierarchy

Van de Poel (2013, p. 262) defines specification as: *'as the translation of a general value into one or more specific design requirements '* and states that this can be done in two steps:

1. Translating a *general value* into one or more *general norms*;
2. Translating these *general norms* into more *specific design requirements*.

For step 1 two criteria are relevant: (1) the norm should be an appropriate response to the value and (2) the norm should be a sufficient response to the value. In step 2 the requirement should be more specific regarding the scope of applicability, goals and aims strived for, and actions to achieve those aims of the norm (Van de Poel, 2013). The value hierarchy has been applied to various cases, for example AI for Social Good (AI4SG) (Umbrello & Van de Poel, 2021) and smart home systems (Umbrello, 2020).

This translation might prove to be quite difficult as insight is needed in the intended use and context of the value which is not always clear from the start of a design project. Also, as artefacts are often used in an unintended way or context, new values are being realized or a lack of values is discovered (van Wynsberghe & Robbins, 2014). An example of this are drones that were initially designed for military purposes, but are now also used by civilians for filming events and even as background lights during the 2017 Super Bowl halftime show. The value of safety is interpreted differently for military users that use drones in desolated regions compared to that of 300 drones flying in formation over football stadium in a populated area. The different context and usage of a drone will lead to a different interpretation of the value *safety* and could lead to more strict distance norms for flight safety which in turn could be further specified in alternate design requirements for rotors and software for proximity alerts, to name two examples.

The application of a value hierarchy to Autonomous Weapon Systems can for example be illustrated by Figure 4 in which the value of *accountability* is translated into norms for `transparency of decision-making' and `insight into the algorithm' (Verdiesen, 2017). This translation will allow users to get an understanding of the decision choices the Autonomous Weapon System makes in order to trace and justify its actions. The norms for *transparency of decision-making* lead to specific design requirements. In this case a feature to *visualise the decision-tree*, but also to *present the decision variables* the Autonomous Weapon Systems used, such as trade-offs in collateral damage percentages of different attack scenarios to provide insight into the proportionality of an attack. The Autonomous Weapon System should also be able to *present the sensor information*, for example imagery of the site, in order to show that it discriminated between combatants and non-combatants. To get *insight into the algorithm*, an Autonomous Weapon System should be designed with features that it normally will not contain. In this case these features would include a *screen* as user interface that shows the algorithm in a *human readable form* and the functionality to *download* the changes made by the algorithm as part of its machine learning abilities that can be studied by an independent party, such as a war tribunal of the United Nations if the legality of the actions of an Autonomous Weapon Systems are questioned.

Kroes and van de Poel (2015) state that an objective measurement of values is not possible because the operationalization is done by means of second-order value judgments which seriously undermine the construct validity of the value measurement. Judgments are often considered subjective as their truth, or falsity, depend on feelings or attitudes of the person who judges (Searle, 1995). To counter this lack of validity, the designer could look to technical codes and standards which are drafted by committees and represent reasonable standards of operationalizing and measuring values in design. However, standards may not reflect the latest technical and social developments and operationalization still requires value judgments of the designer. Kroes and van de Poel (2015, p. 177) advise to *'embed them in a network of other considerations, including definitions of the values at stake in moral philosophy (or the law), existing codes and standards, earlier design experiences, etc.'*.

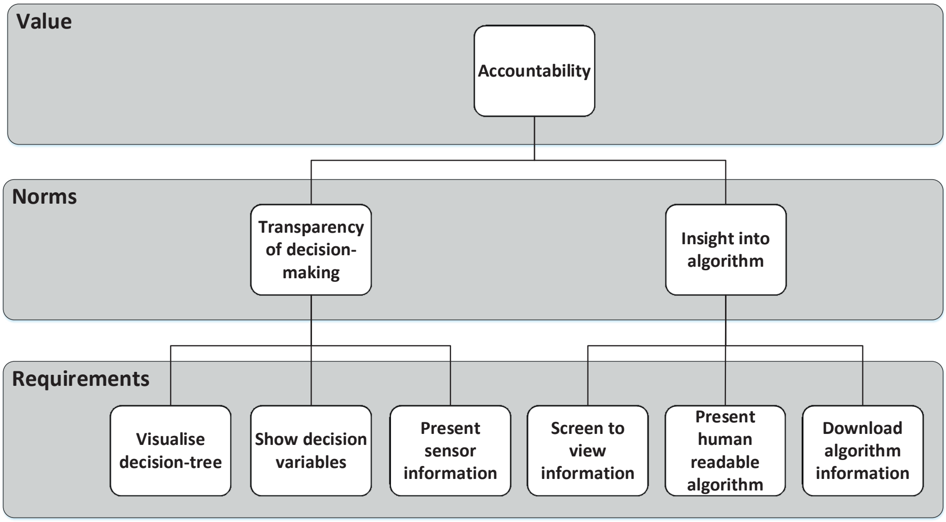

Value
Accountability
Norms
Transparency of decision-making
Insight into algorithm
Requirements
Visualise decision-tree
Show decision variables
Present sensor information
Screen to view information
Present human readable algorithm
Download algorithm information

Figure 4: Value hierarchy for accountability over Autonomous Weapon Systems (Verdiesen, 2017)

In our research, we follow the advice of Kroes and van de Poel (2015) and do not strictly apply the value hierarchy as a method to specify, design and test requirements for Autonomous Weapon Systems. The value hierarchy in Figure 4 is used as orientation, inspiration and direction for our research.

2.8 RESPONSIBILITY

Responsibility can be forward-looking to actions to come and backward-looking to actions that have occurred. Van de Poel (2011) focusses on moral responsibility for consequences to describe the notions of forward- and backward-looking responsibility and does not describe organizational, social and legal responsibility nor responsibility for actions. Two varieties of responsibility that are primarily forward-looking are: 1) responsibility as virtue and 2) the moral obligation that something is the case; and three varieties that are primarily backward-looking are: 3) accountability, 4) blameworthiness and 5) liability.

More formally, forward-looking responsibility is defined by (Van de Poel, 2011, p. 41):

1) *A is forward-looking responsible for X to B means that A owes it to B to see to it that X*

In which A and B are agents (i.e. persons or a forum) and X can be a task, action, outcome or realm of authority. This statement reflects that persons can have specific responsibilities to different people that they owe different responsibilities that might even conflict.

Backward-looking responsibility is formally defined as (Van de Poel, 2011, p. 42):

2) *A is backward-looking responsible for X to B means that it is fitting for B to hold A responsible for X*

This statement entails that being responsible includes being accountable or blameworthy. In this sense accountability performs functions of scrutiny, for example calling someone to account, requiring justifications and imposing sanctions (Mulgan, 2000). The notion of fitting refers to the appropriateness for someone to hold another accountable under certain conditions. The conditions for which it is appropriate or fitting to hold A backward-looking blameworthy are (Van de Poel, 2011):

1. *Capacity condition*: the agent has the capacity to act responsibly i.e. has moral agency;
2. *Causality condition*: the agent is causally connected to the outcome by either an action or an omission;
3. *Wrong-doing condition*: a reasonable suspicion that an agent did something wrong, or could have prevented something wrong from happening and the agent has the burden-of-proof to show that it is not to blame by giving account. The shift of burden-of-proof to the agent that is supposed to have done something wrong only seems reasonable if there are arguments for the suspicion of wrongdoing.

These forms of responsibility are conceptually and casually related in many ways. For instance, one can arguably be deemed to be a responsible person (virtue) only if she accepts blame and liability when needed and is willing to account for his or her actions (G. Williams, 2008). A general capacity for accountability is arguably the basis for other forms of backward-looking responsibility, including blameworthiness (Gardner, 2007). Moral blameworthiness (in the form of culpability or fault) grounds many forms of criminal and tort liability. And by encouraging accountability, it is probably possible to make persons more able and willing to discharge their (forward-looking) moral and social obligations (Pesch, 2015). However, these forms of responsibility are also distinct and require different conditions to apply. For instance, Van de Poel (2011) states that an agent can have backward-looking responsibility (i.e. being accountable or blameworthy) without being forward-looking responsible for preventing that state-of-affairs. Also, blameworthiness requires that an agent has unjustifiably and inexcusably committed a wrong action. Whereas accountability simply requires the agent to explain her behaviour, possibly but not necessarily with the goal of showing that it was not wrong, or that thought wrong, given the circumstances, justifiable or excusable (Gardner, 2007). Also, Pesch (2015) discussed the concept of "active responsibility" of engineers. Active responsibility could be viewed as forward-looking responsibility as it proactively requires engineers to take societal values of technology into account during the development of technology. It is also paired with 'passive' responsibility, also referred to as accountability. The pairing of active responsibility and passive responsibility creates a proactive feedback loop of responsibility that is neither strictly forward-looking nor backward-looking responsibility and by this, it takes an intermediate position between these two types of responsibility. This proactive feedback loop enables actors to learn and reflect on their actions.

Yet another notion is that of command responsibility which originates in the military legal domain and is a concept used in relation to violations of the laws of war and International Human Rights Law (IHRL) (see Table 6 in appendix C for an overview). Command responsibility means that a superior can be held accountable for the crimes committed by his or her subordinates. It originates from the failure of military or civilian superiors preventing their subordinates committing crimes that violate the laws of war or IHRL or not meeting the obligation to punish the violators after committing the crime. Three conditions need to be met for command responsibility to be pertinent: 1) *'The existence of a superior–subordinate relationship was demonstrated by the commander's 'effective control' over the persons who commit the crime, 2) the superior knew or had reason to know that the criminal act was about to be or had been committed, and 3) the superior failed to take the necessary and reasonable measures to prevent the criminal act or punish the perpetrator thereof.'* (Saxon, 2016, p. 24). Command responsibility can be viewed as a combination of virtue-based responsibility and accountability. It is meant as an instrument to hold commanders accountable, but it is also linked to moral identity as the commander has the moral obligation to prevent crimes and violations that are about to be committed.

All forms of responsibility are arguably to be encouraged and promoted in order for Autonomous Weapon Systems to be designed, introduced, regulated and used in a morally acceptable way, and many different forms of responsibility gaps have to be avoided to prevent negative ethical and societal effects of this introduction and use (Santoni de Sio & Mecacci, 2021). However, whereas the relationship between control and blameworthiness has been widely studied in philosophy (Fischer & Ravizza, 1998) and the relation between moral and legal culpability, its gaps, and Meaningful Human Control have been studied in relation to Autonomous Weapon Systems, an account of the relationship between accountability, its gaps, control and oversight is still missing. In the next sections we start filling this lacuna.

2.9 ACCOUNTABILITY

Accountability is a key concept in political science, public management, international relations, social psychology, constitutional law and business administration literature. In the policy domain, the term accountability has two different uses. On the one hand, it is used to praise or criticize the performance of states, organizations, firms or officials regarding policy or decisions in relation to their ability and willingness to give information and explanations about their actions ('accountability as a virtue'). Typically, in the political discourse, accountability is used to describe the fairness and equitability of good governance in which authorities are being held accountable by their citizens. In this broad sense, accountability encompasses concepts such as transparency, equity, democracy, efficiency, responsiveness, responsibility and integrity. On the other hand, in a narrow sense, accountability is also used to define the mechanisms for corporate and public governance to hold agents and organisations accountable ('accountability as a mechanism') (Bovens, Schillemans, & Goodin, 2014). Bovens (2007, p. 450) focuses on the latter sense of accountability and defines it as follows: '*Accountability is a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences*.' The relationship between an actor and a forum is a key notion in the concept of accountability. If the explanation is inadequate, sanctions may be imposed on the actor by a forum (Bovens, 2007; Greer, Wismar, Figueras, & McKee, 2016). Figure 5 provides an overview of the relationship between the accountability elements. Accountability is not only scrutiny after the event has occurred, it also has a preventive and anticipatory use for which norms are (re)produced, internalized and adjusted by means of accountability if necessary.

Similarly, in public administration, mechanisms of accountability are described in terms of an agent having to report on his or her activity to an individual, group or other entity which has the ability to impose costs to the agent (Keohane, 2003). In this sense, accountability is an agency theory approach in which the relationship between a principal and an agent is described (Hulstijn & Burgemeestre, 2014). This concept of accountability as answerability is most used in public administration, but according to Romzek and Dubnick (1987) accountability can play a greater role than answerability alone. It is also linked to the means that public agencies have to manage internal and external expectations of their stakeholders. To manage these internal and external expectations two factors are critical: '*1) whether the ability to define and control expectations is held by some specified entity inside or out-side the agency; and 2) the degree of control that entity is given over defining those agency's expectations*' (Romzek & Dubnick, 1987, p. 228). This notion of accountability is linked to control of expectations of the agency.

Depending on the different relationships between different actors and fora, Bovens (2007) distinguishes five types of (narrow) accountability:

1. *Political accountability* in which the chain of principal-agent relationship, in a democracy being the representatives of voters that form cabinets of ministers, are accountable for the work of public servants;
2. *Legal accountability* is based on specific responsibilities and detail laws and regulations. It is enforced by civil or administrative courts and it is the most unambiguous type of accountability;
3. *Administrative accountability* is enforced by independent external administrative and financial supervision by quasi-legal forum such as auditing offices and (national or local) ombudsmen;
4. *Professional accountability* is based on codes-of-conduct and practices that are created by professional associations, for example in hospitals and schools, and enforced by professional supervisory bodies;
5. *Social accountability* is a recent form of accountability that has been on the rise due to the internet. Non-governmental organizations, interest groups and the public are stakeholders that public organizations feel obliged to give account to regarding their performance by means of public reporting and establishment of public panels. Bovens (2007) notes that this type of accountability might not be seen as a full accountability mechanism because the possibility of judgement and sanctions are lacking, and the relationship between the actor and forum is not clearly described.

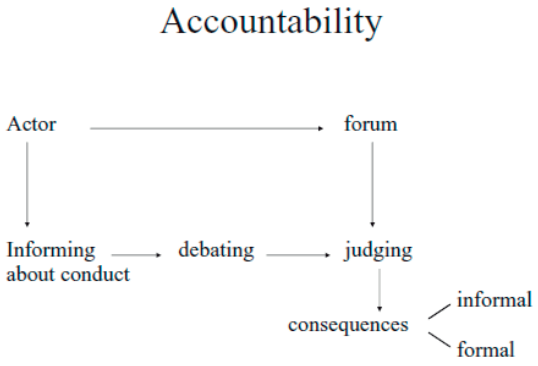


Figure 5: Elements of accountability concept (as in: Bovens, 2007)

Accountability gaps

Many scholars point to accountability gaps that may occur in the deployment of Autonomous Weapon Systems. However, what the authors below refer to as accountability is what we in this research, based on the work of Van de Poel (2011), call ‘blameworthiness’ or ‘culpability’. Those notions are related to backward-looking responsibility, but not similar to the concept of accountability as we employ in this research. In this section we identify these different uses of the term accountability and relate them to the work of Van de Poel (2011). Asaro (2016) argues that the use of emerging technologies, including Autonomous Weapon Systems, with weak or without norms can lead to limited or easily avoidable responsibility and accountability for states and individuals. Sparrow (2016), building on the work of Matthias (2004) and Roff (2013), states that the use of an Autonomous Weapons System might risk an ‘responsibility gap’ and it could be problematic to attribute responsibility for actions taken by Autonomous Weapon Systems to operators. Galliott (2015) also mentions the responsibility gap put forward by Sparrow and argues that shifting to forward-looking responsibility, instead of only backward-looking responsibility, and a functional sense of responsibility to include institutional agents and the human role in engineering the system, might be a solution to avoid this gap. Crootof (2015) also discusses the accountability gap and notes that with the use of Autonomous Weapon Systems serious violations of international humanitarian law may be committed resulting in a lack of criminal liability, which is a form of backward-looking responsibility but not accountability in a strict sense as meant by Van de Poel (2011), for people, including the deployer, programmer, manufacturer and commander, or the weapon system itself. According to Horowitz and Scharre (2015) the potential of an ‘accountability gap’ is the main motivation to implement the principle of Meaningful Human Control. If an Autonomous Weapon System malfunctions and strikes the wrong target it is possible that no human is responsible for the error of the weapon.

Alston (2010) describes these gaps as an 'accountability vacuum' in his UN report to the Human Rights Council on targeted killings. He defines targeted killings as '… *the intentional, premeditated and deliberate use of lethal force, by States or their agents acting under colour of law, or by an organized armed group in armed conflict, against a specific individual who is not in the physical custody of the perpetrator*.' (Alston, 2010, p. 26) notes that states failed to disclose: '…*the procedural and other safeguards in place to ensure that killings are lawful and justified, and the accountability mechanisms that ensure wrongful killings are investigated, prosecuted and punished*.' The reason for this accountability vacuum is that the international community cannot verify the legality of the killing, nor confirm the authenticity of the intelligence used in the targeting process or ensure that the unlawful targeted killing results in impunity. Meloni (2016) argues that the accountability vacuum that Alston described in 2010 has been growing ever since. Cummings (2006a) notes that an erosion of accountability could be caused by the use of computer decision-making systems, because these systems diminish the user's moral agency and responsibility due to the perception that the automated system is in charge. This could cause operators to cognitively offload responsibility for a decision to a computer which can be viewed as a lack of forward-looking (virtue) responsibility. Which in turn creates a moral buffer, meaning a form of distancing and compartmentalizing of decisions, leading to moral and ethical distance and an erosion of accountability.

As we have highlighted above, many authors use different notions when describing accountability gaps. Often, they refer to the notion of accountability, whilst they actually express blameworthiness, culpability or virtue responsibility based on the characterization of Van de Poel (2011). To gain a better understanding of accountability gaps we aim to delineate these gaps in more detail. We identify accountability gaps on three different levels which are based on the layers described by Van den Berg (2015) who distinguishes an engineering, socio-technical and governance perspective to characterize cyberspace. As offloading responsibility of decisions by operators to Autonomous Weapon Systems may lead to erosion of accountability, we identify three possible accountability gaps on three different levels:

1. *Technical accountability gap*: if the system is designed to be technically inaccessible then human operators cannot give a meaningful account of an action mediated by this machine as information on decisions of the machine cannot be retrieved.
2. *Socio-technical accountability gap*: human operators do not have sufficient capacity (skill or knowledge) to interpret the behaviour of the machine even though the behaviour is accessible to, for example, an expert. This is linked to the capacity condition for blameworthiness described by Van de Poel (2011). Also, motivation to interpret the behaviour of a system could be lacking if sufficient mechanisms for accountability are not available.

3. *Governance accountability gap*: an institutional setting is lacking to pressure human operators and other personnel (e.g. commanders, engineers) to account for their (mediated) actions even when the human operator may have the capacity to give a meaningful account. The lacking of an institutional setting also prevents providing protection of the individuals at the lower levels of institutional decisions and omissions.

In the next section we describe the link between accountability and control by following Bovens' (2007) argument that accountability is a form of control, but not all control forms are accountability mechanisms. We characterize control based on an engineering, socio-technical and governance perspective based on the layers described by Van den Berg (2015) (see Figure 6) and briefly highlight where these perspectives fall short. Next, we move to the concept of Meaningful Human Control and argue that social institutional and design dimension at a governance level is needed, because accountability requires strong mechanisms for oversight. We look at an oversight mechanism to connect the technical, socio-technical and governance perspective of control in order to improve accountability for the behaviour of Autonomous Weapon Systems.

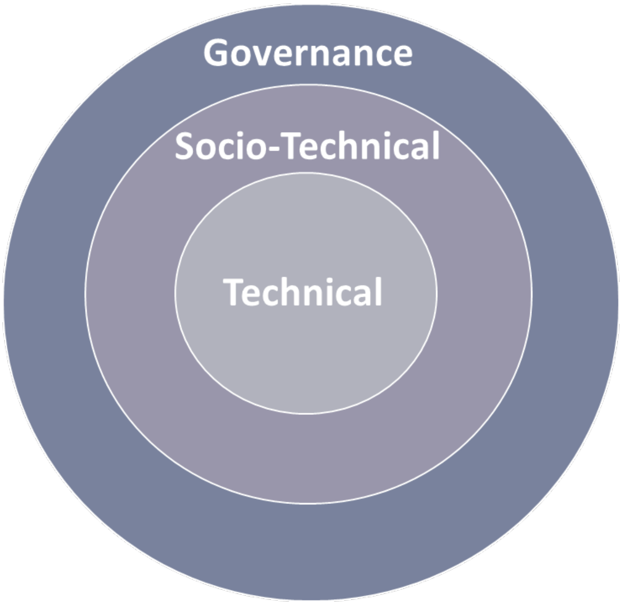


Figure 6: Conceptualization cyberspace in layers (based on Van den Berg, 2015)

2.10 FROM ACCOUNTABILITY VIA CONTROL TO HUMAN OVERSIGHT

Several scholars describe the relationship between accountability and control. According to Bovens (2007) there is a fine line between accountability and control. Koppell (2005, p. 97) states that: *'If X can induce the behavior of Y, it is said that X controls Y—and that Y is accountable to X.'* Radin and Romzek (1996) link types of accountability relationships to the degree (high or low) and source (internal or external) of control. Koppell (2005) notes that this seems to mix different types of accountability relationships which is in his sense a weakness of this approach. According to Lupia (in Bovens 2007, p. 453): '*An agent is accountable to a principal if the principal can exercise control over the agent*'. Bovens (2007) contests this by stating that although accountability mechanisms are important to control the behaviour of organizations, control in the Anglo-Saxon sense means '*having power over*' and can be achieved by '*very proactive means of directing conduct*'. Examples of these proactive means are direct orders, laws, regulations and directives. These means are not accountability mechanisms themselves because they are not procedures in which an actor has to justify and explain his or her conduct to a forum. Bovens (2007) concludes by stating that: '*Accountability is a form of control, but not all forms of control are accountability mechanisms.*'

The question then is if human control can ground effective mechanisms of accountability in relation to the behaviour of agents and institutions who deploy Autonomous Weapon Systems. We will argue, that we need to broaden this view towards oversight, and more specifically what we will call: Comprehensive Human Oversight mechanisms.

2.11 CONTROL

Control has traditionally been defined in different ways, depending on application domains. In this section we describe the perspectives from the engineering, socio-technical and governance point of view based on the layers described by Van den Berg (2015) (see Figure 6).

Engineering perspective

Control from an engineering perspective can be described as a mechanism that compares the output of another system or device to the input and goal function by means of a feedback loop to take action to minimize the difference between outcome and goal. These control systems can range from very simple, e.g. household thermostats, to very complex, for example nuclear power plant control (Åström & Kumar, 2014; Pigeau & McCann, 2002). In general, a control system has four common characteristics: (1)

it is a *dynamic system* with responses that evolve in time and has memory of past responses, (2) it requires *stability* to function without failure, (3) it contains a *feedback mechanism* with sensors and detectors to determine the accuracy of control, and (4) *dynamic compensation* to approximate the performance limits of the components of the control system (Kheir et al., 1996). The traditional engineering perspective holds a very mechanical or cybernetic view on the notion of control, one that is not well-suited to make sense of the interaction between a human agent and an intelligent system for which the human is to remain accountable.

Socio-technical perspective

The socio-technical perspective on control describes which agent has the power to influence the behaviour of another agent (Koppell, 2005). An agent can be human or a technological system. The influence of one agent over another is often mediated by technology and it also includes controlling the technology. It involves instruments to direct the behaviour of agents like legal regulations, sanctions or political instructions (Mulgan, 2000). Unlike the engineering one, this notion of control is intrinsically connected to the achievement of shared (social) tasks and goals, concerns the relation between human agents and it is therefore potentially relevant to the idea of accountability. Scott (2000) makes a distinction between *ex ante* and *ex post* control. *Ex ante* involvement in decision-making is related to managerial control and accountability-based control is linked to *ex post* oversight. Busuioc (2007) also conceptualizes control based on this temporal dimension. She differentiates three types of control in a principal-agent relationship:

1. *Ex ante or proactive control* which is a preliminary control mechanism that defines the boundaries of the autonomy of agents to achieve a delegated task;
2. *Ongoing or simultaneous control* which is an informal type of direct control of an agent that specifies the goals but not the specific actions an agent has to take to achieve a delegated task;
3. *Ex post control or accountability* which is the principle of delegating powers to an agent and therefore renounced direct control. It is a process of providing information, discussion and evaluation to determine the extent to which the agent has lived up to its *ex ante* mandate and has acted within its zone of discretion after the fact.

Control from a socio-technical perspective is power-oriented and aimed to influence behaviour of agents making use of *ex ante*, *ongoing* or *ex post* instruments. However, it does not explicitly include mechanisms of power over nonhuman intelligent systems, like Autonomous Weapon Systems.

Governance perspective

The governance perspective on control describes which institutions or forums supervise the behaviour of agents to govern their activities. Pesch (2015) argues that there is no institutional structure for engineers which calls on them to recognize, reflect upon and actively integrate values into the designs on a structural basis. The result is that the moral effects of a design can only be evaluated and adjusted after the implementation in society. Pesch (2015) notes that engineers relate to different institutional domains, such as the market, the state and science. The consequence is that engineers do not have a clearly defined accountability forum and that they rely on engineering ethics and codes of conduct. However, these codes of conduct are often not robustly enough institutionalized to be regarded as a good regulative framework. Therefore, engineers use methods such as the Value-Sensitive Design and Constructive Technology Assessment as proxies for accountability forums. The need to develop and use these proxies for engineering practices reveals that a governance perspective on responsibility and control lacks robust institutionalized frameworks.

The insufficiency of traditional notions of control to make sense of the human control over Autonomous Weapon Systems required to ground accountability, has led to the introduction of the notion of Meaningful Human Control in the political debate on Autonomous Weapon Systems. However, a common definition of this notion has been lacking in practice for a long time (Ekelhof, 2019). Some scholars have been working on defining and operationalizing Meaningful Human Control over the past years. Horowitz and Scharre (2015, pp. 14-15) were one of the first to list three essential components for Meaningful Human Control: '(1) *Human operators are making informed, conscious decisions about the use of weapons.* (2) *Human operators have sufficient information to ensure the lawfulness of the action they are taking, given what they know about the target, the weapon, and the context for action.* (3) *The weapon is designed and tested, and human operators are properly trained, to ensure effective control over the use of the weapon.*' However, these three components do not apply to Autonomous Weapon Systems alone, but apply to the use of weapons in general. Ekelhof (2019) states that the relationship between the human operator and Autonomous Weapon System is used as reference to define Meaningful Human Control, but this is still a general and abstract definition of this notion. Moreover, this notion of control has a very operational view and is strongly, if not exclusively, focused on the relation between one human controller and one technical system, and tries to identify the different conditions under which that controller may be able to effectively interact with the system. We may call this a narrow notion of Meaningful Human Control, insofar as the broader perspective of governance of control, organisational aspects, values and norms does not seem to be incorporated.

In an attempt to overcome the conceptual impasse on the notion of Meaningful Human Control, Santoni de Sio and Van den Hoven (2018) tried to offer a deeper philosophical analysis of the concept, by connecting it more directly to some concepts coming from the philosophical debate on free will and moral responsibility, and in particular the concept of "guidance control" by Fischer and Ravizza (1998). By reinterpreting and adapting the two criteria for guidance control, they eventually identified two conditions that need to be satisfied for an autonomous system to be under Meaningful Human Control. The first condition is the *tracking* condition that entails that '*the system should be able to respond to both the relevant moral reasons of the humans designing and deploying the system and the relevant facts in the environment in which the system operates…*'. The second condition is the *tracing* condition according to which the actions of an Autonomous (Weapon) System should be traceable to a proper technical and moral understanding on the part of one or more relevant human person who designs or interacts with the system (Santoni de Sio & Van den Hoven, 2018, p. 1).

Mecacci and Santoni De Sio (2019) operationalized this concept of Meaningful Human Control even further in order to specify design requirements. They focused on the tracking condition and offer a framework for which Meaningful Human Control as "reason-responsiveness" which identifies agents and their different type of reasons in relation to the behaviour of an automated system. By this, Mecacci and Santoni De Sio (2019) go beyond engineering and human factors conceptions of control. In a way that directly connects Meaningful Human Control with the idea of social control over the technology, the authors reason that, in presence of appropriate technical and institutional design, a system can and should be under Meaningful Human Control by more than one agent and even by super-individual agents such as a company, society or state. These complex relationships of "reason-responsiveness" are modelled in a framework that looks at the distance of different forms of human reasoning to the behaviour of a system. This scale of distance allows for classifying different type of agents and their contexts, values and norms. Mecacci and Santoni De Sio's (2019) framework shows that the narrow focus of engineering and human factors control needs to be widened to allow a development of autonomous technologies that are sufficiently responsive to ethical and societal needs. In recent years, other scholars have been working on operationalising the concept of Meaningful Human Control (see section 7.1. for emerging insights on operationalising Meaningful Human Control). Amoroso and Tamburrini (2021) created a normative framework for Meaningful Human Control. They suggest a differentiated approach and to abandon the search for a one-size-fits all solution. They state that rules are needed to bridge the gap between specific weapon systems and their uses on one hand and the ethical and legal principles on the other hand. Another approach is that of Umbrello (2021) in which he couples two different Levels of Abstraction (LoA) to achieve Meaningful Human Control over an Autonomous Weapon System. In this, he combines systems thinking and systems engineering as conceptual tools to frame the commonalities between these two LoAs. A third approach to operationalise Meaningful Human Control is presented by Cavalcante Siebert et al. (2023) who are proposing four actional properties for AI-based systems under Meaningful Human Control to bridge the gap between philosophical theory and engineering practice. In their reflection on their work the authors highlight that '*Meaningful human control is necessary but not sufficient for ethical AI*.' (Cavalcante Siebert et al., 2023, p. 252). The authors amplify this by stating that for a human-AI system to align with societal values and norms, Meaningful Human Control must entail a larger set design objectives which can be achieved by transdisciplinary practices.

However, the wider conception of the control loop mentioned above does not incorporate the social institutional and design dimension at a governance level. The governance level 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. Therefore, we look at an oversight mechanism to connect the technical, socio-technical and governance perspective of control which may ensure solid controllability and accountability for the behaviour of Autonomous Weapon Systems.

2.12 HUMAN OVERSIGHT

Several scholars mention that an oversight mechanism is needed in order to hold an actor accountable (Caparini, 2004; Schedler, 1999; Scott, 2000). West and Cooper (1989: in (Pelizzo, Stapenhurst, & Olson, 2006)) mention two reasons for oversight in the political system: (1) it can improve the quality of policies or programs and (2) when policies are ratified by the legislative branch, they obtain more legitimacy. The oversight mechanism can be implemented as an *ex post* review process or a mechanism for either *ex post* of *ex ante* supervision (Pelizzo et al., 2006).

According to Goodin (1995) responsibility needs supervisory action in that A has to see to it that X is achieved. He states that '*… require[s] certain activities of a self-supervisory nature from A. The standard form of responsibility is that A see to it that X. It is not enough that X occurs. A must also have "seen to it" that X occurs. "Seeing to it that X" requires, minimally, that A satisfy himself that there is some process (mechanism or activity) at work whereby X will be brought about; that A check from time to time to make sure that that process is still at work, and is performing as expected; and that A* *take steps as necessary to alter or replace processes that no longer seem likely to bring about X.*' (Goodin, 1995, p. 83). Supervision has to be done by the agent and cannot be delegated.

Oversight over international institutions can be used as an equivalent for the accountability of these institutions according to De Wet (2008). She distinguishes three forms of oversight: (1) *vertical oversight* in which there is a hierarchy between institutions and the parent organ can exercise formal control over and issue sanctions to the child organ, (2) *horizontal oversight* which is not based on a hierarchical supervisory organ but often is on voluntarily or based on a constitutive document and sanctioning is mostly restricted to social pressure or public naming-and-shaming, and (3) *intermediate oversight*, which lies in between vertical and horizontal oversight and has a formal basis in a constitutive document but is supervised by a non-hierarchical institution which often acts and reports to a body higher up in hierarchy and sanctions vary in severity.

2.13 CONCLUSION

Ethical concerns on Autonomous Weapon Systems call for a process of human oversight to ensure accountability over targeting decisions and the use of force. Responsibility, accountability and Meaningful Human Control are values often mentioned in the societal and academic debate. Ongoing control or direct control (Busuioc, 2007) by an (human) agent is not possible in case of executive autonomy because the notion of executive autonomy described by Castelfranchi and Falcone (2003) has implications for the applicability of (military) control instruments for Autonomous Weapon Systems. Bovens (2007) notes that accountability can be viewed as a form of control, but not all forms of control are accountability mechanisms. Similarly, Meaningful Human Control, at least in Santoni de Sio and Van den Hoven (2018) perspective, not always requires more traditional forms of technical control such as direct power of a human controller, or a competent human operator having a constant and meaningful interaction with the technical system, even though these may sometimes be needed. But accountability always 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. Therefore, based on the literature review above, we propose a Framework for Comprehensive Human Oversight that connects the engineering, socio-technical and governance perspective of control. By this we broaden the view on the control over Autonomous Weapon Systems and take a comprehensive approach that goes beyond the notions of control described above. In the next chapter, the Comprehensive Human Oversight Framework is elucidated and applied to the military domain.

3|

Conceptual Framework

In the previous chapter we state that accountability is a form of control and the notion of control can be viewed from different perspectives. In this chapter we follow this line of reasoning and present the Comprehensive Human Oversight Framework that contains different control mechanisms to ensure accountability over Autonomous Weapon Systems. By describing the layers and the connections between them we identify two gaps in the control mechanisms. To mitigate these gaps, we applied the Glass Box framework (a framework for monitoring abstract values and translating them into observable elements) on the Comprehensive Human Oversight Framework (a framework that depicts control and governance mechanisms - see section 3.1). We conclude by closing the loop from the review stage back to the interpretation stage by means of a feedback process. Parts of this chapter have been published in Verdiesen, De Sio, and Dignum (2019) and Verdiesen, Aler Tubella, and Dignum (2021).

3.1 COMPREHENSIVE HUMAN OVERSIGHT FRAMEWORK

In the literature review of the conceptual investigation phase of our research, we studied the work of Van den Berg (2015) and his three-layered model consisting of a technical, socio-technical and governance layer (see section 2.9) that he created to describe cyber space. In our analysis we linked these layers to the accountability mechanisms (section 2.9) and control perspectives (section 2.11) to a time perspective which shows when a process is taking place. This analysis led to the design of the Comprehensive Human Oversight Framework (Figure 7).

On the x-axis of the Comprehensive Human Oversight Framework *time* is plotted which can be divided into three phases: (1) before deployment of a weapon, (2) during deployment of a weapon and (3) after deployment of a weapon. These phases are depicted by the vertical columns of the framework. The y-axis describes the *environment* of the system which can range from more internal to more external to the technical system. The combination of layers and columns result in nine blocks that each contain a component of control in each phase and layer. For example, before deployment the *input* to a system is a component to control the goal of the system in the technical layer. The Comprehensive Human Oversight Framework allows to highlight the existence of gaps in control. These are presented below in section 3.3. Figure 7 depicts the three layers of the Comprehensive Human Oversight Framework. The bottom technical layer describes the internal environment of the system and the upper governance layer the external environment of the system. The middle socio-technical layer is the intersection between the internal and external environment.

Technical layer
The technical layer describes the technical conditions required for the system to remain under control. The system should be able to receive the right input (for example restrictions on the boundaries of the area of operation) from the human operator (block 7), the system's feedback mechanism should be robustly and verifiably to check the difference between output and goals during development (block 8) in order to keep responding to the reasons (goals and norms) of the human operators, and after deployment it should be technically possible to verify and understand the output (e.g. check if the system did not cross the boundaries of the allocated geographical area) and the processes behind them (block 9).

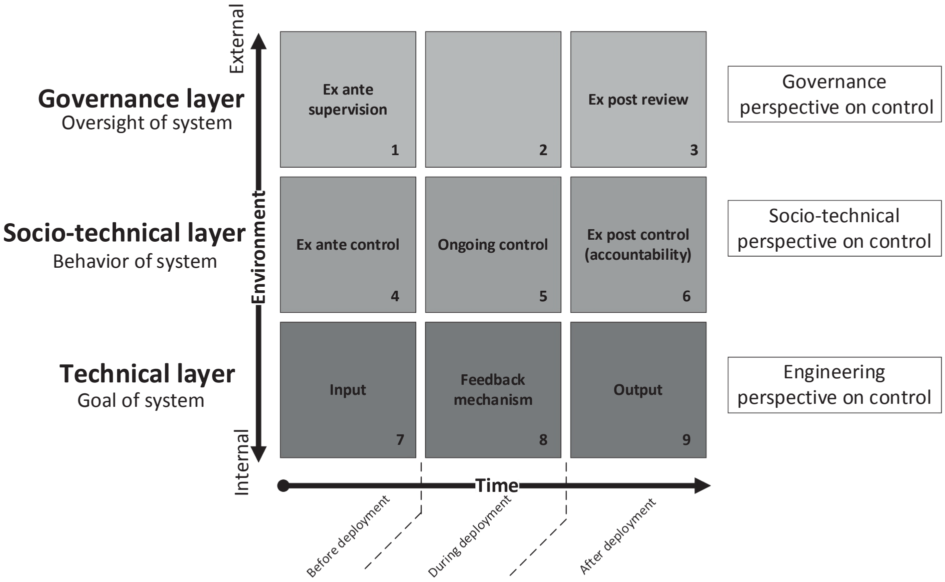


Figure 7: Comprehensive Human Oversight Framework

Socio-technical layer

The socio-technical layer describes the operators' psychological and motivational conditions required for the system to remain under control. Ex ante, the human operators should be able to set the right control measures before deployment and to correctly appreciate the capabilities and limitations of the systems (block 4). *Ex ante control* is a preliminary control mechanism that defines the boundaries of the autonomy of agents to achieve a delegated task (Busuioc, 2007). During use, the human operators should have the capacity to have a meaningful interaction with the system and understand what it is doing in order to supervise the system to have ongoing control (block 5). *Ongoing control* is an informal type of direct control of an agent that specifies the goals but not the specific actions an agent has to take to achieve a delegated task (Busuioc, 2007). Supervision is seeing to it that something is achieved by an actor (see section 2.12). Ex post, after deployment of the system, the human operators should be able to inspect and assess the behaviour of the system to be able to account for its actions (block 6). *Ex post control* is the principle of delegating powers to an agent and therefore renouncing direct control. It is a process of providing information, discussion and evaluation to determine the extent to which the agent has lived up to its ex ante mandate (Busuioc, 2007).

Governance layer

The governance layer describes the political and institutional conditions and the oversight mechanisms required for the system to remain under control. Before deployment institutional and political mechanisms, such as fora, clear definitions of the roles of accountor and accountee, should be put in place to exert ex ante supervision (block 1). After deployment an ex post review process ensures that the fora have the power to demand an account and sanction if the account is not satisfactory (block 3). As far as the literature study found, there is no process to oversee the system during deployment (block 2). The oversight of the system in the governance layer is conducted before and after deployment by the ex-ante supervision and ex-post review processes, but an oversight mechanism during deployment seems to be lacking.

Both the horizontal layers and vertical columns are interconnected and depend on each other for information. For example, without appropriate input to a system in the technology layer (block 7), there is no feedback loop (block 8) and output (block 9). The output of the technology layer (block 9) is in turn needed to be able to account for as ex post control mechanism (block 6) in the socio-technical layer. This accountability mechanism of block 6 feeds into the ex post review process (block 3) of the governance layer. The components clearly also have causal interconnections. Most notably, the presence (or lack thereof) of adequate ex-ante governance mechanisms (block 1) would affect all the other components, all the way to the technical output of the system (block 9). Also, any gap in these connections will cause problems at the lower levels.

In Figure 7 a clear gap is visible in the governance layer of the middle column. Based on our literature study a mechanism in block 2 appears to be missing indicating a gap in the governance layer. As an oversight process seems to be lacking, there is no sufficient mechanism for an institution to govern or supervise the ongoing control (block 5) of a system in the socio-technical layer. The lack of an oversight mechanism in block 2 may lead to deficiencies in the ongoing control mechanism in block 5. In turn this influences the ex post control or accountability mechanism in block 6 as there is no instrument, mechanism or process for an institution in the accountability process to confirm if the conduct during the deployment of the weapon, for which should be accounted for in a forum, actually occurred as there is no monitoring process of an independent institution during deployment. This in turn could lead to deficiencies in the ex post review process (block 3) of the governance layer and could impede both the active responsibility during deployment as the backward-looking responsibility after deployment.

Validation

When drafting the Comprehensive Human Oversight Framework, we actively sought feedback on the design from academic, military and industry experts during presentations and group discussions. During the empirical investigation phase of this research the aim of the interviews we conducted was to discuss the Comprehensive Human Oversight

Framework and its alignment with the Design for Values approach. This feedback allowed us to improve the Comprehensive Human Oversight Framework iteratively. However, for academic rigor the Framework should be validated and evaluated in a study to verify if it holds and to evaluate it. Future work should focus on validating the Comphrensive Human Oversight Framework with a structured scientific method to review and improve it if necessary.

The next section presents the Dutch military control instruments that are currently used in the layers and the weapon deployment phases of the Comprehensive Human Oversight Framework. In applying the Comprehensive Human Oversight Framework to the military domain, we identified a process used in military operations that fills the void in the governance layer during deployment. We will describe this process more in detail in the next section. Subsequently, we describe the connections and feedback loop between the layers. We conclude by recommending to close the feedback loop in the governance layer to incorporate the findings of the review process in the mandate for a next mission.

3.2 APPLICATION OF THE COMPREHENSIVE HUMAN OVERSIGHT FRAMEWORK TO EXISTING MILITARY CONTROL INSTRUMENTS

From a military perspective, control is described as a process to check if current and planned orders are on track and if the objectives to achieve a goal are met (Alberts & Hayes, 2006; Liao, 2008; NATO, 2017). Control aims to make adjustments to the plan if the current state deviates from the planned end-state of the mission. Control measures bound the mission space by limiting the area of operation, duration of military operations and by defining the order of battle. Control consists of procedures for planning, directing and coordination of resources for a mission and this includes standard operating procedures (SOPs), rules of engagement (ROEs), regulations, military law, organizational structures and policies (Pigeau & McCann, 2002). Control in a military perspective is an instrument to bound and check if the actions are in line with the planned military goal and to adjust the planning when the current state deviates from the end state. This resembles the notion of control in an engineering perspective because there is a goal, input and feedback loop to adjust the system.

In the military domain a variety of instruments are used as control mechanisms before, during and after deployment of weapons in military operations. After our analysis of the control mechanisms in the governance, socio-technical and technical perspectives on control in section 2.9, we turned to the military domain to identify the military control instruments that are currently used in the three layers. We identified that in the military domain there is a control mechanism in each layer before, during and after deployment of a weapon system. We found that the targeting process is used in military operations to plan and direct activities during deployment of a weapon. The targeting process is defined as a process that '*links strategic-level direction and guidance with tactical targeting activities through the operational-level targeting cycle in a focused and systemic manner to create specific physical and psychological effects to reach military objectives and the desired end state.*' (Ekelhof, 2018, p. 66). Based on our analysis we plotted the military control mechanisms in the Netherlands on the Comprehensive Human Oversight Framework (Figure 8). These mechanisms are described below.

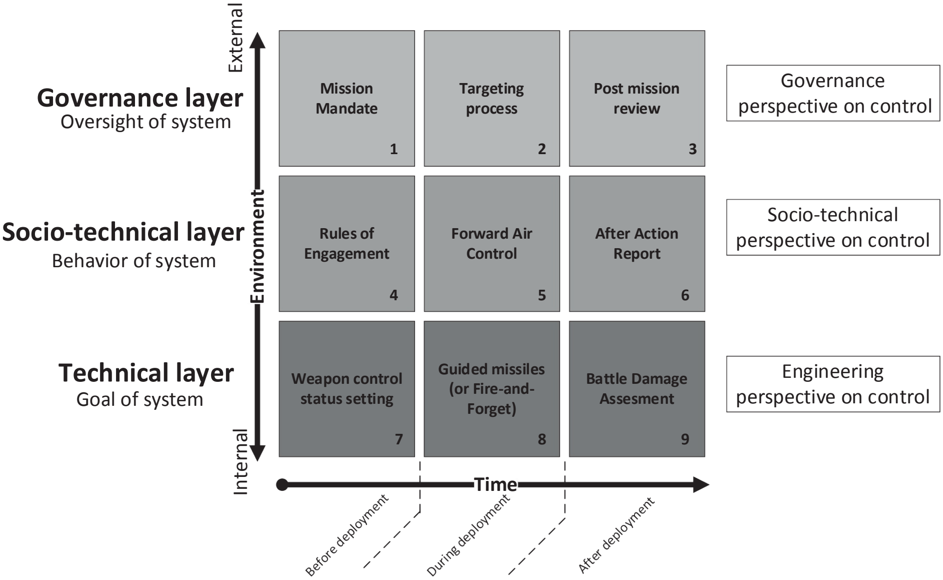


Figure 8: Military control instruments plotted on the Comprehensive Human Oversight Framework

In the text below we describe the existing military control instruments that we have plotted on to our Comprehensive Human Oversight Framework. The military control instruments in the Comprehensive Human Oversight Framework per block as defined for The Netherlands armed forces are:

1. *Ex ante supervision*
   Before a mission a Mission Mandate is issued by the UN or NATO. This instrument is the result of political consideration and describes the tasks of a specific mission before troops are deployed. It does not contain specificities on weapon deployment.

2. *Targeting process*

   During deployment the targeting process is a deliberate iterative decision-making cycle for methodical planning of actions to counter opponents in order to achieve the effect in the strategic and operational campaign plan. The targeting process consists of six phases: (1) commander's intent, objectives and guidance, (2) target development, (3) capabilities analysis, (4) commander's decision, force planning and assignment, (5) mission planning and force execution and 6) assessment (Ekelhof, 2018).

3. *Ex post review*

   In the Netherlands, after a mission is finished, it is evaluated to inform parliament on the results and progress of the mission. The evaluation report is published online and mentions Rules of Engagement and number of weapon deployments. In some cases, the government decides to conduct a post mission review 5 years after a mission as a second evaluation. This is only done when asked for by the government and is not a structural process.

4. *Ex ante control measures*

   Several control instruments are used before deployment to control the usage of weapons. These are amongst others the Rules of Engagement, assignment of command relationships and determining the Area of Responsibility (AOR).

5. *Ongoing control measures*

   During a mission the deployment of a weapon can be done by a Forward Air Controller who can employ different levels of control to release a weapon.

6. *Ex post control measures*

   In the Netherlands, an After Action Report (AAR) is filed after each weapon deployment which is send via the Military Police to the Public Prosecution Office of the Department of Justice.

7. *Input*

   The instrument used to control weapons before deployment, is the Weapon Control Status Setting in which the level of control of a weapon is determined after a deliberation process.

8. *Feedback*

   Some weapons, e.g. guided missiles, have a feedback loop and can be controlled during launch, but most weapons are fire-and-forget systems that do not have a feedback loop once launched.

9. *Output*

   The output of weapon deployment is the destruction of a target in order to achieve a military effect. A Battle Damage Assessment (BDA) is conducted to assess if the effect is achieved and to assess the (collateral) damage inflicted on a military objective.

Contrary to the analysis of the academic literature describing the control mechanisms in the governance, socio-technical and technical perspective in Figure 7, the military domain has an oversight mechanism during deployment in block 2 (see Figure 8). The targeting process in block 2 is a decision-making process for methodical planning of actions to counter opponents in order to achieve the effect in the strategic and operational campaign plan (NATO, 2016). The targeting process is a domain specific process for the military and is not monitored by an independent institution. By this, it is comparable to the statement of Pesch (2015) that an institutional structure for engineers is lacking to call on them to recognize, reflect upon and actively integrate values into the designs on a structural basis. Like engineers, the military does not have an independent institutional structure to call on them to reflect upon their values and principles during deployment. Reflection is done within the military domain and if military personnel violate military law and regulations they have to account for their conduct at a military court. However, this accountability process will be conducted after deployment and is not part of the targeting process during deployment.

The military control instruments in Figure 8 are connected in the vertical columns of the layers. For example, the Rules-of-Engagement (block 4) will be based upon the Mission Mandate (block 1) and the options for weapon control status setting (block 7) will be determined by the Rules of Engagement (block 4). This is also the case for the horizontal levels as the Rules of Engagement (block 4) determine the guidelines of the Forward Air Control (block 5) and the After Action Report (block 6). This also applies to the bottom-up process after deployment. The Battle Damage Assessment (block 9) will be input for the After Action Report (block 6). The After Action Reports (block 6) should be used in the Post Mission Review process (block 3).

The feedback loop in the governance level from the Post Mission Review process (block 3) to the Mission Mandate (block 1) is often not conducted. A reason for this might be that different institutions are responsible for these instruments. The UN or NATO will draft the Mission Mandate and the Post Mission Review process is a national instrument. It is difficult to embed a national perspective in a multilateral document. In the socio-technical and technical level this feedback loop is conducted more often as these are within the military sphere of influence. For example, the Rules of Engagement (block 4) can be adjusted based on the findings of After Action Reports (block 6) and the Forward Air Control procedures (block 3) can be changed in accordance with the Rules Of Engagement (block 4). We recommend to try to close the feedback loop in the governance level so that findings in the Post Mission Review process will feed back into the Mission Mandate.

In the next section we apply the Comprehensive Human Oversight Framework to the case of Autonomous Weapon Systems. We describe the implications for the applicability of military control instruments for Autonomous Weapon Systems with different levels of autonomy. We compare the Comprehensive Human Oversight Framework presented in section 3.1, which is based on the literature review, to the application of the Framework to Autonomous Weapon Systems (section 3.3). This reveals two gaps in the control mechanisms that arise when the concept of autonomy is introduced in weapon systems which can be linked to the accountability gaps in section 2.9.

3.3 APPLICATION COMPREHENSIVE HUMAN OVERSIGHT FRAMEWORK TO AUTONOMOUS WEAPON SYSTEMS

The difference between a conventional weapon system and an Autonomous Weapon System is the notion of autonomy (see section 2.3). Weapon systems may comprise of different levels of autonomy. But even in the case of a "fully Autonomous Weapon System", '*[…] that, without human intervention, selects and engages targets matching certain predefined criteria, following a human decision to deploy the weapon on the understanding that an attack, once launched, cannot be stopped by human intervention.*' (AIV & CAVV, 2015; Broeks et al., 2021) the type of autonomy can at most be executive autonomy (see section 2.3 for description of executive autonomy), because a human will set its goals and the weapon will not decide on its goals or deployment itself. Also, the context will constrain the autonomy of a "fully Autonomous Weapon System" as autonomous systems are created with task goals and boundary conditions (Bradshaw, Hoffman, Woods, & Johnson, 2013). In case of Autonomous Weapon Systems, the context will include physical limitations to the area of operations, for example the presence, or lack of, civilians in the land, sea, cyber, air or space domain.

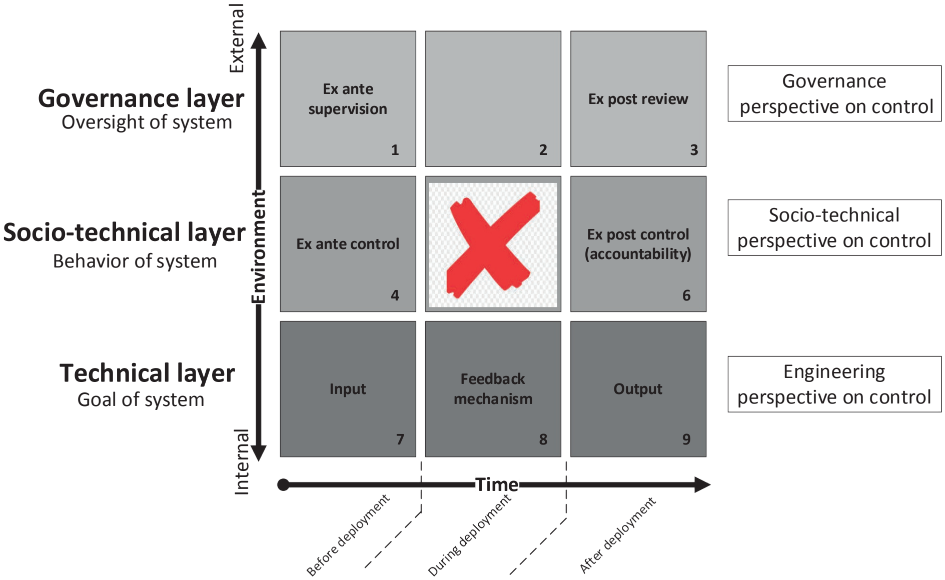


Figure 9: Comprehensive Human Oversight Framework for Autonomous Weapon Systems

This notion of executive autonomy has implications for the applicability of military control instruments for Weapon Systems with different levels of autonomy, including fully Autonomous Weapon Systems. In the different phases executive autonomy implies that:

a. *Before* deployment of an Autonomous Weapon System,
   i. In the technical layer the human will set the input (e.g. predefined criteria),
   ii. This will be based on the ex-ante control measures, for example the Rules of Engagement, in the social-technical layer.
   iii. and this will be done within the boundaries of the ex-ante supervision mechanism, such as the mission mandate, in the governance layer.

b. *During* deployment of an Autonomous Weapon System,
   i. In the technical layer, the Autonomous Weapon System itself conducts the feedback loop, as found in most (industrial) control systems, to take action to minimize the difference between outcome and goal (for example heat seeking missiles).
   ii. In the socio-technical layer the mechanism of ongoing control means that the goals are specified by a human before deployment, but the human does not specify the actions that the weapon has to take to achieve that goal. There is no ongoing control mechanism or instrument for fully Autonomous Weapon

Systems to control these specific actions that the Autonomous Weapon System takes to achieve its goal, because executive autonomy inherently implies that the main subject (*x*) (i.e. the Autonomous Weapon System) is autonomous in setting its means (i.e. actions) to achieve its goal (μ) independently from secondary subject (γ) (i.e. the human operator). Partially Autonomous Weapon Systems may be designed to respond to the input of operators or controller, but given the complexity and speed of these systems, it is an open question to what extent and under which conditions operators and controllers would be able to effectively supervise and intervene (see section 2.11 on Meaningful Human Control).

iii. In the governance layer an independent mechanism to monitor these actions of the Autonomous Weapon System is missing in the current Comprehensive Human Oversight Framework (see Figure 9).

c. *After* deployment of an Autonomous Weapon System,

i. The output of weapon deployment in the technical layer is the destruction of a target in order to achieve a military effect and the output will be verified by a Battle Damage Assessment (BDA),

ii. There is an ex-post control mechanism to account for the weapon deployment in socio-technical layer, being the After Action Report (AAR) process.

iii. The ex post review in the governance layer could be done to evaluate the mission in a post mission review process and takes the Rules of Engagement and number of weapon deployments into account.

The current military control mechanisms described above are sufficient to bound the area of operation, the duration of the operation and deployment of weapons. But the introduction of autonomy in Autonomous Weapon Systems has implications on the military control mechanisms, mainly in the socio-technical layer during deployment of an Autonomous Weapon System. This may require reformation of the military control instruments. These implications might lead to new training methods for military personnel for them to have the capacity (knowledge and skills) to responsibly deploy these weapons, but might also lead to new institutions and design methods, for example Value-Sensitive Design in (military) engineering (Santoni de Sio & Van den Hoven, 2018), as control mechanisms in the governance layer.

Comparing the Comprehensive Human Oversight Framework in Figure 7 to that of Autonomous Weapon Systems in Figure 9 reveals two gaps in the control mechanisms that can be linked to the accountability gaps in section 2.9: (1) a mechanism in block 2 of an independent institution that ensures oversight of a weapon during deployment (a governance accountability gap), and (2) in the Comprehensive Human Oversight

Framework for Autonomous Weapon Systems there is no ongoing control mechanism in block 5 to control the specific actions that the Autonomous Weapon System takes to achieve its goal (a socio-technical accountability gap). On the one hand, fully executive autonomy inherently implies that the Autonomous Weapon System is autonomous in setting its means to achieve its goal independently from the human operator. On the other hand, even less-than fully Autonomous Weapon Systems may still present big challenges in allowing the human controller to have effective control and supervision. This may actually depend, among other things, on the extent to which the ex ante and ex post mechanisms of control over the human–machine interaction are sufficient to give the operator the relevant capacities and motivation to discharge her duties. At a broader level, this arguably also depends on the extent that the governance level can provide an acceptable level of control on the choice of weapons and the distribution of tasks and duties in the mission.

To mitigate the governance and socio-technical accountability gaps, we applied the Glass Box framework - a framework for monitoring abstract values and translating them into observable elements - on the Comprehensive Human Oversight Framework. In section 3.4 we will describe the Glass Box framework before applying it to the Comprehensive Human Oversight Framework in section 3.5.

3.4 GLASS BOX FRAMEWORK

As stated in section 3.1, based on our literature study, a mechanism in block 2 of the Comphrensive Human Oversight Framework appears to be missing indicating a gap in the governance layer. As an oversight process seems to be lacking, there is no sufficient mechanism for an institution to govern or supervise the ongoing control (block 5) of a system in the socio-technical layer. Next to this, introducing the notion of executive autonomy has implications for the applicability of military control instruments for Weapon Systems with different levels of autonomy, including fully Autonomous Weapon Systems (see section 3.3). This means that there is no ongoing control mechanism or instrument for fully Autonomous Weapon Systems to control these specific actions that the Autonomous Weapon System takes to achieve its goal during the deployment phase. To fill these gaps a mechanism is needed to monitor the compliance of norms to ensure accountability over autonomous systems. The Glass Box framework could serve as a mechanism to solve these gaps, because it monitors abstract values and translates them into observable elements. The Glass Box approach (Aler Tubella, Theodorou, Dignum, & Dignum, 2019) is a framework (see Figure 10) for monitoring adherence to the contextual interpretations of abstract values which focuses uniquely on the observable inputs and outputs of an intelligent system. Its focus on the observable aspects of the system's behaviour makes it particularly apt for monitoring autonomous and generally opaque systems.

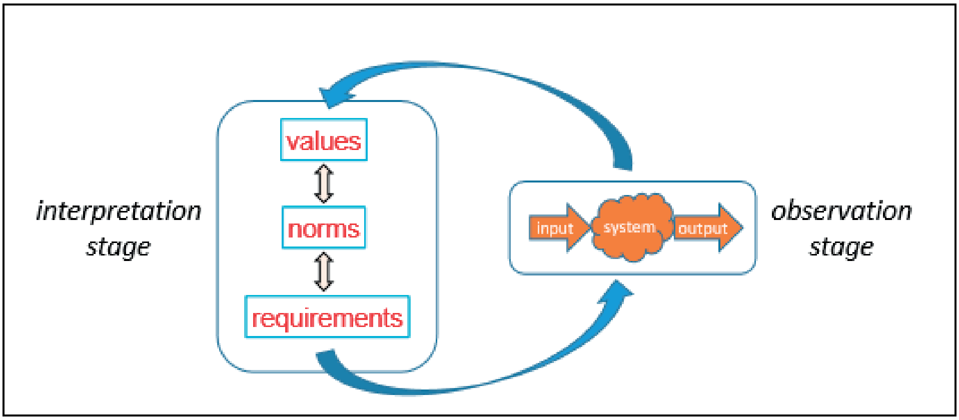


Figure 10: Glass Box framework (as in: (Aler Tubella et al., 2019))

The Glass Box approach consists of two phases which inform each other: interpretation and observation. The interpretation stage consists of a progressive process of concretising abstract values into specific design requirements. Following a *Design for Values perspective* (Van de Poel, 2013), the translation from values to requirements is done by considering the different stakeholder interpretations and contexts. The output from the interpretation stage is an abstract-to-concrete hierarchy of norms where the highest level is made up of values and the lowest level is composed of fine-grained concrete requirements for the intelligent system only related to its inputs and outputs. The intermediate levels are composed of progressively more abstract norms, where fulfilling a concrete norm "counts as" fulfilling the more abstract one in a certain context. This hierarchy of norms transparently displays how values are operationalised, together with which contexts have been considered.

The second phase of the approach is given by the observation stage. This stage is informed by the requirements on inputs and outputs identified in the interpretation stage, as they determine what must be verified and checked. In the observation stage, the system is evaluated by studying its compliance with the requirements identified in the previous stage: for each requirement, we assign one or several tests to verify whether it is being fulfilled. The difficulty of these tests can range from an extremely simple yes/no check on whether an action has been performed, to sophisticated statistical analysis depending on the type of norms identified.

Feedback between interpretation and observation stage throughout the lifespan of the system is necessary: continuous observation informs us on which requirements are consistently unfulfilled, which may prompt changes in the implementation or in the chosen requirements. This approach therefore transparently monitors and exposes possible malfunctions or misuse of the system.

Ensuring accountability and adherence to values in the context of drone or Autonomous Weapon System deployment is inextricably tied to the notion of human oversight and human accountability. For this reason, we propose to consider drone and Autonomous Weapon System deployment a "process within a socio-technical system", the monitoring of which includes not only examining the behaviour of the drone or Autonomous Weapon System itself but also examining human-led procedures in pre- and post-deployment. A specific adaptation of the Glass Box approach to this context is therefore the explicit inclusion of the operator(s) as an entity to which norms can apply.

A significant choice in this framework is the decision to consider the drone and Autonomous Weapon Systems a "black box", the internal logic of which is not accessible. This responds to two motivations. Firstly, relying on access and monitoring capabilities on the internal workings of drones and Autonomous Weapon Systems would be a strong assumption, since the proprietary nature of this technology often precludes observation of its software. Second, for auditability purposes, the users of this framework should be able to transparently follow the monitoring process. However, such users, who will respond to the monitoring process, do not necessarily possess the technical background required to understand or check constraints on the internal logic of a drone. Thus, our framework is based on monitoring adherence to norms constraining purely observable elements of pre-, and post-deployment of the drone or Autonomous Weapon System. Another choice is that we purposely designed a technology-agnostic approach so that it can be used on many different systems independent from the AI techniques and algorithms that are used as internal workings of the drone or Autonomous Weapon System. We consider these as part of the black box.

In what follows, we present an adaptation of the Glass Box approach for the inclusion of human oversight in autonomous drone or weapon deployment. The proposed framework includes an interpretation and an observation stage, each discussed in detail.

Interpretation stage

The interpretation stage entails turning values into concrete norms constraining observable elements and actions within the socio-technical system in a similar way as is done in the value hierarchy (Van de Poel, 2013) which is described in section 2.7. This hierarchical structure of values, norms and design requirements makes the value judgements, that are required for the translation, explicit, transparent and debatable. As high-level concepts, values are abstract, whereas norms are prescriptive and impose or forbid courses of action. Such a translation is done by constructing norms progressively, subsuming each norm into several more concrete ones, until the level of norms containing concrete testable requirements is reached. This concretisation of norms will be carried out by all stakeholders involved in the deployment, ideally with legal advisory as well as with participation from operators themselves (whose processes will be subject to the norms identified).

Through a *Design for Values* perspective (Cummings, 2006; Davis & Nathan, 2015; Friedman, Kahn Jr, Borning, & Huldtgren, 2013; Van de Poel, 2013; van den Hoven, Vermaas, & van de Poel, 2015), concretising values requires carefully adapting to the specific context, as values may take different meanings in different contexts. In the case of drone and Autonomous Weapon System deployment, the context is made up of two main factors: the context of deployment itself, and the organisation doing the deployment. Thus, some norms may generally apply to any deployment (such as organisational rules), whereas others may be highly specific (such as regulations governing specific areas or purposes). For this reason, the interpretation stage does not produce a one-size-fits-all normative framework, but rather it needs to be updated in any change of context. The specific tying of norms to a context enforces human oversight in this stage: new human-designed norms are needed for any new context of deployment, thus necessarily implicating the deploying organisation in the process of considering each situation's specificity and risk.

Even though values and their interpretations vary by culture, purpose, organisation, and context, some values are fundamentally tied to the context of drone deployment. As with any technology deployed into society, a fundamental value is that of lawfulness. A requirement for any drone or Autonomous Weapon System deployment is, for example, to respect flight rules (e.g., maximum height of flight and avoidance of airport surroundings). Thus, the identification of requirements for the trajectory taken by the drone or Autonomous Weapon System is a fundamental aspect of this stage. Given the different capabilities that drones or Autonomous Weapon Systems may be equipped with, aspects of the law related to flying over public spaces, commercial liability, or privacy (Rao, Gopi, & Maione, 2016), as well as surveillance (Rosén, 2014) or warfare, must be considered. The purpose of deployment itself (e.g., humanitarian aid, commercial delivery, or bird observation) will determine the relevant values that guide the process, such as privacy (Luppicini & So, 2016), safety (Clarke & Moses, 2014), humanity (van Wynsberghe & Comes, 2020), or ecological sustainability (Vas, Lescroël, Duriez, Boguszewski, & Grémillet, 2015).

Design requirements need to refer to the observable behaviour of drone or Autonomous Weapon System and operator, and are considered in the context of pre- and post-flight procedures. They may apply to checkable behaviours of the drone or Autonomous Weapon System (flying over a certain altitude or flying over certain areas), to pre-flight processes (getting approval or checking weather conditions), or to post-flight processes (evaluation of route followed or treatment of the data obtained). Crucially, they are not limited to the drones' or Autonomous Weapon Systems' behaviour, but must include the system around it for human oversight: procedures such as pre-flight safety checks, acquiring authorisations or human review of the data obtained should all be mandated and constrained, so that we can guarantee that the entire flight process has been subject to human oversight. The norms and observable requirements identified at this stage form the basis for the next stage, indicating what should be monitored and checked, and which actions constitute norm violations.

Observation stage

In this stage, the behaviour of the system is evaluated with respect to the values by studying its compliance with the requirements identified in the interpretation stage. As these requirements focus on observable behaviours, in this stage observations are made, and it is reported whether norms are being adhered to or not (and, by extension, whether values are being fulfilled).

Observations can be automated (e.g., automatically trigger a flag if the drone or Autonomous Weapon System has deviated from its planned path), or manually performed by an operator, depending on the requirement. A specific trade-off to consider is the observation time versus the reliability of the observations: extensive, lengthy manual or computationally expensive checks may take a long time to perform, delaying operations, but may be the only way to check a certain requirement. Depending on how crucial such a requirement is, observations may be relaxed (e.g., performed at random intervals), or the requirement modified for a better fit.

From these observations, we can compute whether norms have been adhered to. Such a computation can be done through a formal representation of the norms and requirements. For example, a formalisation of the Glass Box can be found in Aler Tubella and Dignum (2019), using a "counts-as" operator to relate more concrete norms to their more abstract counterparts. Within that formalisation, by assigning ground truth values to a set of propositional atoms through the observations, we can compute which norms have been adhered to, and escalate up the hierarchy of norms to determine which values have been followed in each context. Alternatively, norms can, for example, be expressed in a deontological language (Wright, 1981) and similarly relate to the observations by representing them as ground truths. A different, complementary approach that we describe in the chapter 5 is the use of Coloured Petri Nets (CPNs) as modelling language for the requirements. By adding tokens to different states depending on the observations (roughly, adding a token if the observation is positive, and not if it is negative), we can simulate the pre- and post-flight processes and determine whether it proceeds correctly or whether norm violations have occurred.

The outcome of the observation stage is either a confirmation that all specifications have been followed, or evidence of norm violations given by the observations that trigger the violation. Human oversight requires that such violations entail accountability processes and a review of the process culminating in the “failed” flight. By providing concrete evidence of where such failures to follow the specifications occurred, this framework therefore explicitly enables oversight without requiring access to the internal logic of the machine, ensuring accountability.

3.5 COMPREHENSIVE HUMAN OVERSIGHT FRAMEWORK PROJECTED ON GLASS BOX FRAMEWORK

When the two stages of the Glass Box framework are projected on the Comprehensive Human Oversight Framework, Figure 11 is generated. The Interpretation stage of the Glass Box framework, in which values in the governance layer are turned into concrete norms, constraining observable elements and actions in the socio-technical layer, which in turn are translated into requirements in the technical layer, is done before deployment—visible in the first column of Figure 11.

During deployment the behaviour and actions of an autonomous system are monitored in the governance layer and verified in the technical layer in the Observation stage of the Glass Box framework that treats the block in the socio-technical layer as a black box visible in the middle column of Figure 11. After deployment a Review stage is required as an accountability process in which a forum in the governance layer can hold an actor in the socio-technical layer accountable for its conduct in the technical layer—visible in the third column of Figure 11. The outcome of the Review stage should feed back into the Interpretation stage for a next deployment of an autonomous system and thereby close the loop between the stages.

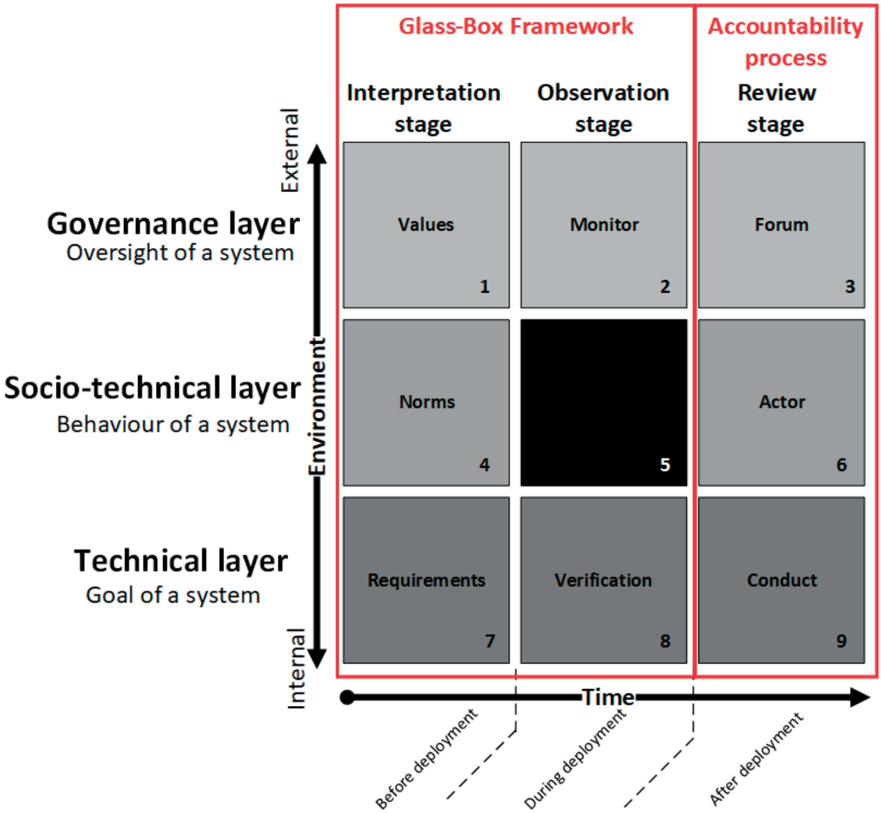


Figure 11: Glass Box framework projected on Comprehensive Human Oversight Framework

3.6 FEEDBACK LOOP: CLOSING THE GAP

Our (previous) research covers the green arrows (see Figure 12) from value elicitation (chapter 4), deriving norms and requirements (chapter 5), monitoring and verification of the norms (chapter 5). Research by other scholars is done on deriving conduct from verification in the field of explainable AI (XAI) (black arrow [1]). Bovens (2007) describes the accountability process form conduct, actor and forum (black arrow [2]) (see section 2.9). The arrow from the accountability process back to the interpretation stage is still a gap that needs to be filled to close the feedback loop. In this section literature on this feedback loop (red arrow [3]) is discussed and a framework for closing the loop is applied to the Comprehensive Human Oversight Framework.

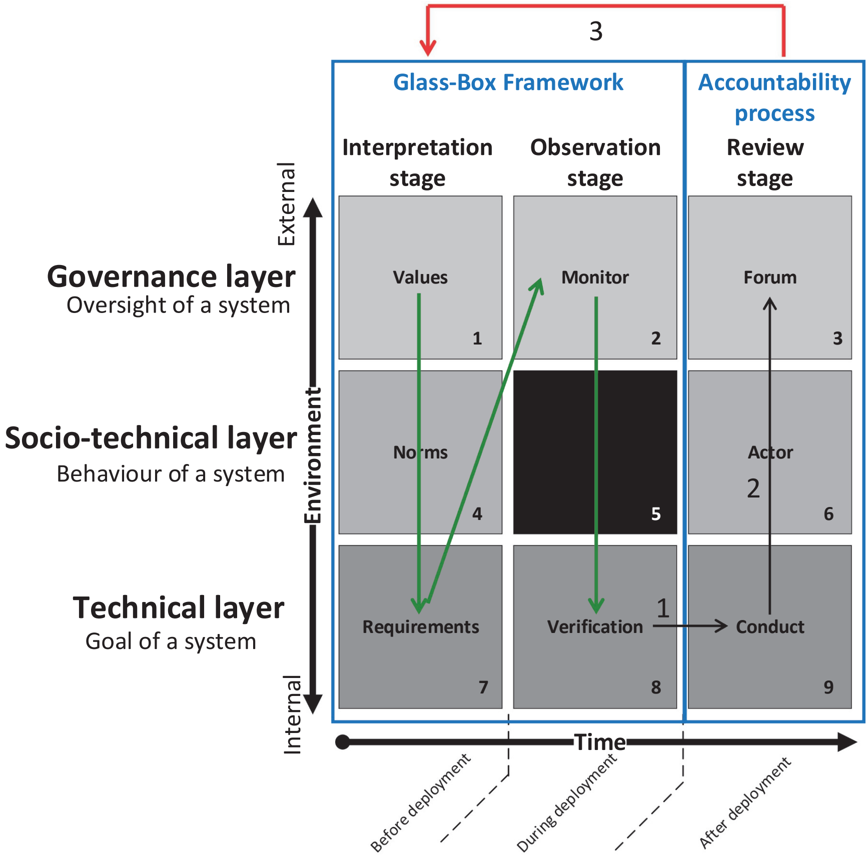


Figure 12: Feedback loop Comprehensive Human Oversight Framework

Application Five-Point Systems Framework on Comprehensive Human Oversight Framework

Accountability provides external feedback on intended and unintended effects of policies. It can stimulate learning, reflecting and improving performance using a feedback mechanism (Bovens, 2014). According to Jacobs (2010) a feedback system can create an institutional link between participatory processes and management systems. Feedback systems can raise significant ethical issues that mirror concerns in participatory practice. The most significant barrier to implement feedback systems appears to be the incentives that shape management and organizational behaviour (Jacobs, 2010). A feedback loop is described as "*a systematic approach to collecting the views of [beneficiaries] and other key stakeholders about the quality and impact of work undertaken by a development agency.*" (Gigler et al., 2014, pp. 212-213).

Three interconnected steps are identified in a feedback loop: 1) sharing information, 2) giving feedback, and 3) taking action and communicate back. These steps can be achieved by applying a Five-Point Systems Framework described by Gigler et al. (2014) that holds five components: *purpose, people, process, tools* and *environment*. *Purpose* describes the broader ends that feedback tries to facilitate. It is a critical component for a feedback system, it shapes performance expectations for those providing, responding and evaluating the feedback so that the architecture of the feedback systems facilitates the objectives. The *people* component relates to choosing who can participate. Selecting participants is a trade-off between inclusivity and complexity. It should identify the roles and responsibilities of all stakeholders within the feedback loop. This includes not only the people providing feedback, but also considering who is monitoring, responding and acting on feedback. The selection of involved actors can have socio-political implications and might alter the power dynamics of stakeholders. *Process* is about developing rules and norms for engaging those who provide feedback. The process should describe the type and frequency of feedback, how it will be integrated and the organizational capacity that is needed to manage the feedback mechanism. Choosing the right tools will help to expand reach and ensure inclusiveness. *Tool*s can be no tech, low tech and high tech depending on the environment and people that need to be reached. *Environment* encompasses the formal and informal societal norms that can increase the inclusiveness of the process. One of the greatest challenges is catalysing and sustaining motivation to participate in feedback mechanisms. Creating an inclusive environment can help to prevent participation fatigue which might occur when participation is not reflected in the final policy or product. Closing the feedback loop requires an organizational effort and capacity in order to implement sustainable, inclusive and efficient feedback (Gigler et al., 2014).

In this section the Five-Point Systems Framework is applied to the Comprehensive Human Oversight Framework to describe the five components needed to close the feedback loop (red arrow [3] in Figure 12). We conclude this section with recommendations for further research to validate the feedback loop.

1. <u>Purpose</u>
   The purpose of the feedback system from the accountability process during the review stage of the Comprehensive Human Oversight Framework is to ensure that the lessons and recommendations from the review stage will be incorporated in the interpretation stage before deployment of an Autonomous Weapon System in a next iteration. The feedback can be incorporated in the elicitation of values in the governance layer (block 2 of Figure 12), the derivation of norms from these values in the socio-technical layer (block 2 of Figure 12) or requirements in the technical layer for an autonomous system (block 3 of Figure 12).

2. People

Two questions need to be answered when considering who is involved and is allowed to participate in the feedback. The selection of involved actors can have socio-political implications and might alter the power dynamics of stakeholders (Cornwall 2008; Mohan 2001 in: Gigler et al., 2014)

A. *Who Provides the Feedback?*

In the accountability process during the review stage the *forum* that holds an *actor* accountable for its *conduct* is the entity that provides feedback. Depending on the type of accountability, political, legal, administrative, professional or social accountability (Bovens, 2007), a different forum provides the feedback. In case of *political* accountability, it is the politicians of political parties who, as representatives of voters, are the forum. For *legal* accountability, civil or administrative courts are the forum that holds actors accountable. *Administrative* accountability is enforced by quasi-legal forums such as auditing offices and (national or local) ombudsmen. *Professional* accountability is based on codes-of-conduct and practices that are created by professional associations, for example in hospitals and schools, and enforced by professional supervisory bodies as forums. Finally, *social* accountability is a recent form of accountability in which non-governmental organizations, interest groups and the public are stakeholders that public organizations feel obliged to give account to regarding their performance by means of public reporting and establishment of public panels.

In the case of a feedback mechanism for the Comprehensive Human Oversight Framework a forum could be a political entity in case of a formal investigation to the conduct (or responsibility) of a minister for political accountability. Legal accountability in case of investigation, the forum is a criminal court or a trial for war crimes. For professional accountability, a professional supervisory body within a Ministry of Defense could act as a forum. Administrative accountability can be performed by a third-party auditing committee. Social accountability is not applicable as feedback system of a Comprehensive Human Oversight Framework as this is not a formal process and the possibility of judgement and sanctions is lacking.

B. *Who Monitors, Responds to, and Acts on the Feedback?*

The outcome of the review process depends on the type of accountability that is applicable. In case of:

- political accountability, it could be a report with directions based on a hearing, inquires or proceedings process;

- legal accountability, the outcome of a trial or court ruling is a decision or court order;
- administrative accountability, it could be an investigation report with recommendations;
- professional accountability, could be recommendations, advice or lessons learned.

The review process could lead to recommendations or obligations to incorporate in the governance, socio-technical or technical layer of the next iteration of the interpretation stage of the Comprehensive Human Oversight Framework.

The *monitoring*, *responding* and *acting* on the outcome of the review process can be done by an article 36 weapon review committee. Several countries, at least Belgium, Germany, the Netherlands, New Zealand, Norway, Sweden, Switzerland, the United Kingdom and the United States (Verbruggen & Boulanin, 2017), have an article 36 reviewing procedure in place to conduct weapon reviews when new weapons and methods or means of warfare are studied, developed, acquired or adopted. The aim of the article 36 weapon review is to monitor the development of weapons by reference to its obligations under International Humanitarian Law by a State (McClelland, 2003). However, very few countries have a formal review mechanism in place (Verbruggen & Boulanin, 2017) and the format and responsibilities of the reviewing authority, how states interpret the terms of reference and legal obligations of Article 36 are conducted differ by each country. For example, the United States describes a separate approval process for fully Autonomous Weapons by a senior review committee in their updated DOD DIRECTIVE 3000.09 AUTONOMY IN WEAPON SYSTEMS (US Department of Defense, 2023). Whilst The Netherlands has established an Advisory Commission on International Law and Conventional Weapons Use (AIRCW) which uses a three-step process for an Article 36 review in which the actual review is conducted by a working group (Verbruggen & Boulanin, 2017).

A standard review process is lacking and developing an international standard article 36 review process could ensure the incorporation of the obligations and recommendations in the next iteration of the interpretation stage of the Comprehensive Human Oversight Framework.

3. ***<u>Process</u>***

What type of feedback with what frequency is required is the question that needs to be answered to describe the process. According to Gigler et al. (2014) feedback should be viewed as a typology of types of information or interaction and not be viewed as a monolithic concept. They suggest four types of feedback: complaint, suggestions, monitoring and satisfaction. These types do not fit the case of Autonomous Weapon Systems, because these are based on user or customer feedback. Also, the more feedback you seek the more capacity it takes to respond and act on it. Therefore, the type of feedback should be limited.

In case of the Comprehensive Human Oversight Framework the relevant feedback is information on compliance or non-compliance of the system with the criteria that were set pre-deployment. These criteria should be set during the interpretation stage and can be done during drafting of the norms after the value deliberation.

The frequency of the feedback is dependent on the type of accountability process that is followed:

- Political accountability process will often be conducted as part of the post mission review process or after an incident that requires a political hearing;
- Legal accountability process will be conducted after a violation of the law has occurred;
- Administrative accountability process will be performed on request of an institution or actor;
- Professional account process will be conducted when internal regulation or codes of conduct are violated. In the case of deployment of an Autonomous Weapon System this could be when during the After Action Review a deviation of the pre-determined criteria is observed.

Based on the type of accountability, it is not possible to set a specific pre-determined time frame for the feedback process as it is dependent on the type of accountability process described above. Nevertheless, when an article 36 weapon review process can be standardized as suggested above a frequency can be determined. The article 36 review process is mostly conducted when new weapons and methods or means of warfare are studied, developed, acquired or adopted. In the case of deployment of Autonomous Weapon Systems this frequency needs to be increased due to the autonomous decision-making of AI. Determining the nature of the frequency should be part of the design of the standardized article 36 review process.

4. *Tools*

In selecting the type of tools – no, low or high tech – (see Table 7) for conducting the feedback process two criteria are important; 1) expanding reach by levering new technologies and 2) ensuring inclusivity of participation in order not to reinforce existing inequities. Applying this to the case of the Comprehensive Human Oversight Framework, the people providing feedback, either as part of political, legal or professional accountability or taking part in the article 36 review committee, have access to high tech ICT systems. High tech ICT tools can strengthen the criteria mentioned above by extending the reach of the feedback process and ensuring inclusivity, but this should be purposely considered when designing and implementing high tech ICT systems for a feedback process.

Table 7: Spectrum of ICTs (as in: Gigler et al., 2014, p. 230)

Technology category	*Description and barriers to access*	*Example*
No tech	Relies on in-person interactions; negligible barriers to access[a]	In-person site visits, interviews, community meetings
Low tech	Increasingly ubiquitous and rapidly approaching complete penetration; low barriers to access[a]	Community radio or television, mobile phones (straddles low, high)
High tech	Comparatively new with lower penetration rates; higher barriers to access[a]	Internet, social media, mobile phones (straddles low, high)
In terms of cost, literacy, and hardware.		

Environment

The environment consists of the institutional and cultural context in which formal and informal societal norms guide the interaction in the feedback process. Two obligations are important when designing a feedback process; 1) creating a measure to track the representativeness for those providing feedback and 2) balancing costs and benefits for those who participate. Applying these obligations to the case of the Comprehensive Human Oversight Framework, when designing the feedback process careful consideration on tracking the representativeness and cost-benefit balance is needed. This means that in designing the committees for the various types of accountability mechanisms or in standardizing the article 36 review process, the participants should be inclusive ensuring a wide variety of stakeholder participation, not only governmental stakeholders, but also including non-obvious or vulnerable stakeholders, such as representatives of non-governmental organizations, industry or citizen advocacy organizations. The 'transaction costs' of participating should be balanced with benefits for participating. This balance should be considered when designing the feedback process.

Application Five-Point Systems Framework to Autonomous Weapon System case using a toy example

To apply the Five-Point Systems Framework to the case of Autonomous Weapon Systems, the scenario described in section 1.2 is used as an example.

> *An Autonomous Weapon System provides force protection for soldiers that are clearing the road from improvised explosive devices. The Autonomous Weapon System is equipped with surveillance equipment, weapons (air-to-ground missiles) and flies autonomously in the Area of Operation. It is programmed to avoid flying over a restricted operating zone and an electronic warfare threat. The Autonomous Weapon System is equipped with facial and image recognition software for people, weapons and explosives. It is programmed with different options to engage when it recognizes a threat to the soldiers that are clearing the road. The Autonomous Weapon System detects movement behind a large rock near a narrow part of the road at a distance of 300 meters of the road clearance soldiers.*

The action that the Autonomous Weapon System takes is the following:

> *The Autonomous Weapon System positively identifies with a confidence of 99% using facial recognition all three persons sitting behind the rock as members of an opponent group aiming weapons at the road clearing soldiers. The Autonomous Weapon System automatically engages to neutralize the threat to the road clearance soldiers.*

During the After Action Review (AAR), in which the Battle Damage Assessment information is investigated by the Ministry of Defense, it is discovered that the facial recognition software has identified with a confidence level of 97% that one of the persons was a member of an opponent group and the other two persons were identified with a 100% confidence level. Although the criterium of a confidence of 99% as an average for the group is reached, for one of the individuals the 99% confidence level was not reached. An investigation of the technical requirements of the AWS by a third-party auditing committee revealed that the average confidence level was calculated as input criteria in the observation stage of the AWS and not the individual confidence level.

This finding triggers the Five-Point Systems feedback loop to review this deviation of the input criteria:

1) *Purpose*: The purpose of the feedback system from the accountability process during the review stage of the Comprehensive Human Oversight Framework is to ensure that this finding during the review stage is incorporated in the interpretation stage before deployment of the Autonomous Weapon System in a next iteration. As the investigation of the third-party auditing committee turned out this concerned the requirement of the Autonomous Weapon System at the technical level.
2) *People*: The people providing the feedback are part of a supervisory body that

conduct the AAR, which is a form of professional accountability, and a third-party auditing committee, which is a form of administrative accountability. *Monitoring*, *responding* and *acting* is done by an article 36 review committee after the updated technical requirement(s) and modifications are implemented in the system.

3) *Process*: the process that is followed is the internal AAR process based on the Battle Damage Assessment, the investigation process of the third-party auditing committee, the change process for the technical modification of the Autonomous Weapon System and the article 36 review process to assess the Autonomous Weapon System before it is deployed again.
4) *Tools*: High tech ICT tools are used for the investigation, auditing, technical modification and reviewing of the findings.
5) *Environment*: in this case the stakeholders are the Ministry of Defense, the auditing committee and the industry providing the software and Autonomous Weapon System.

This toy example shows that the Five-Point Systems feedback loop can be applied to the case of an Autonomous Weapon System. Depending on the finding for which the feedback loop is conducted, different people, process and environment are relevant and need to be considered as part of the feedback system.

Validation

The Five-Point Systems Framework described by Gigler et al. (2014, p. 219) is based on lessons learned from their literature review and World Bank practice. Both qualitative and quantitative research was conducted to ground their Five-Point Systems Framework within the context of current practices at the World Bank (Gigler et al., 2014, p. 236) and recommendations are given for future technology-enabled citizen feedback initiatives (Gigler et al., 2014, p. 260-264).

Despite the qualitative and quantitative research to ground the Five-Point Systems Framework by Gigler et al. (2014), this is the first time that the framework has been applied to close the feedback loop of the Comprehensive Human Oversight Framework and applied to the case of Autonomous Weapon Systems. Although merely being a toy example described above, it seems that the Five-Point Systems feedback loop can be applied to the case of an Autonomous Weapon System. However, for academic rigor the Five-Point Systems Framework applied to the Comprehensive Human Oversight Framework and the case of Autonomous Weapon Systems should be validated and evaluated in future work to verify if it holds and to evaluate it.

3.7 CONCLUSION

Accountability is a form of control and the notion of control can be viewed from different perspectives. In this chapter we describe the engineering perspective, the socio-technical perspective and the governance perspective. Our main claim is that combining the control mechanisms in the technical, socio-technical and governance layer will lead to Comprehensive Human Oversight over Autonomous Weapon Systems which may ensure solid controllability and accountability for the behaviour of Autonomous Weapon Systems. These three perspectives on control constitute the three layers of our proposed Comprehensive Human Oversight Framework. The Comprehensive Human Oversight Framework highlights the connection between the layers and shows an existing gap in the governance layer. Current military control instruments cover the blocks of the Comprehensive Human Oversight Framework. However, when applied to the case of Autonomous Weapon Systems the Comprehensive Human Oversight Framework reveals two gaps in control, one gap in the governance layer and one in the socio-technical layer during deployment of an Autonomous Weapon System. The application of the Glass Box framework on the Comprehensive Human Oversight Framework could mitigate these gaps in control.

The Glass Box framework is built around the black box (the autonomous drone or weapon system) with the Interpretation and the Observation stage which allows for a transparent human oversight process which ensures accountability for the deployment of an autonomous system. As this is a first attempt to implement the Glass Box framework in a practical manner further research is needed to validate the concept.

A feedback process can close the loop from the accountability process after deployment of a weapon back to the interpretation stage before a next deployment of a weapon. The *monitoring*, *responding* and *acting* on the outcome of the review process can be done by an article 36 weapon review committee. Sharing best practices of weapon reviews (Sayler, 2021) could be beneficial to improve the feedback process. Nonetheless, a standard process is lacking and developing an international standard article 36 review process could ensure incorporating the obligations and recommendations in the next iteration of the interpretation stage of the Comprehensive Human Oversight Framework. The toy example described above shows that the Five-Point Systems feedback loop (Gigler et al., 2014) can be applied to the case of Autonomous Weapon System. The purpose of the feedback system from the accountability process during the review stage of the Comprehensive Human Oversight Framework is to ensure that the lessons and recommendations from the review stage will be incorporated in the interpretation stage before deployment of an Autonomous Weapon System in a next iteration.

Part III

EMPIRICAL INVESTIGATION PHASE

During the empirical investigation phase, we build on our conceptual work by conducting qualitative research through interviews, value deliberation in expert panels and a survey. This allowed us on the one hand to discuss and reflect on the results from the conceptual investigation phase and on the other to elicit values that are related to the deployment of Autonomous Weapon Systems. The value elicitation provides insight into which values are deemed important in the deployment of Autonomous Weapon Systems. During the interpretation stage of the Glass Box framework, norms and requirements can be derived based on this value elicitation.

4|

Value deliberation

In this chapter we describe the empirical investigation phase of our research which consists of conducting expert interviews, the Value Deliberation Process as a means to elicitate values and validating the results by consulting experts. For reflection and validation, we discussed the Comphrensive Human Oversight Framework and aspects of drone deployments during interviews. The Value Deliberation Process allowed us to elicit values to be used in the interpretation stage of the Glass Box framework. To substantiate the Value Deliberation results an extra round of validation was conducted by inviting experts- who had not been part of the expert panel- to reflect on the findings. Parts of this chapter have been published in Verdiesen and Dignum (2022).

4.1 EXPERT INTERVIEWS

Three expert interviews were conducted to get more empirical background information. We interviewed a professor at Delft University of Technology to check the academic relevance of the Comprehensive Human Oversight Framework. To gain understanding in the empirical context of current drone deployment we interviewed an applied researcher at the NLR [Netherlands Airspace Centre] (previously working in the Royal Netherlands Airforce) and two operators at the drone squadron of the Royal Netherlands Airforce. The questions regarding the drone deployment were on the decision-making processes, mission planning, execution and evaluation of current drone missions and the results were used to create the implementation concept in chapter 5.

4.2 VALUE DELIBERATION PROCESS

For the value elicitation of the interpretation stage of the Glass Box framework we used the Value Deliberation Process developed by (Pigmans, 2020). Value deliberation is a form of participative deliberation aimed at creating mutual understanding on the various perspectives of the participants. By discussing values instead of solutions, a common ground and normative meta-consensus among stakeholders can be achieved (Dryzek & Niemeyer, 2006). Active participation in a debate offers the opportunity for people to develop and draft collective judgements on complex issues in real time. Deliberation will enhance critical thinking and reflection among its participants through a formalized and guided process. Through (online) deliberation, one can find solutions that consider and integrate various views on certain aspects of a topic. It enables people to learn about the different aspects of a complex (political) topic and to better understand each other's positions (Verdiesen, Dignum, & Hoven, 2018). Based on the practical implementation of deliberative democracy platforms, Fishkin (2009) identifies five characteristics essential for legitimate deliberation: 1) *information*: accurate and relevant data is made available to all participants, 2) *substantive balance*: different positions are compared based on their supporting evidence, 3) *diversity*: all major positions relevant to the matter at hand and held by the public are considered, 4) *conscientiousness*: participants sincerely weigh all arguments, 5) *equal consideration*: views are weighed based on evidence, not on who is advocating a particular view. The Value Deliberation process that Pigmans (2020) developed is inspired by the Delphi method. Where the Delphi method is designed to reach consensus between anonymous experts in an iterative process, the Value Deliberation process is aimed at reaching mutual understanding on the various stakeholder perspectives by direct interaction. The Value Deliberation process consists of six stages and eight steps (see Figure 13).

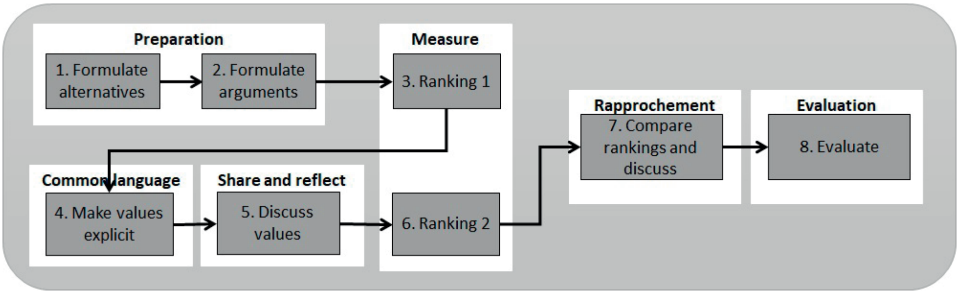


Figure 13: Phases of Value Deliberation Process (adapted from: Pigmans, 2020)

Preparation is phase 1 in which the initiator briefs the topic and if applicable, the predefined solutions to the problem. Next, an independent facilitator takes over and starts with two preparatory steps in conjunction with the participants: step 1- formulate alternatives and step 2 – formulate arguments. Phase 2 consists of measuring by ranking the alternatives from most preferable to least preferable (step 3). A Borda count is used to calculate the individual rankings. In phase 3, a common language is created by the elicitation of values (step 4). These values are discussed in phase 4 to create a mutual understanding (step 5). After a second ranking in step 6- based on the same principles in step 3 – the rankings are discussed and compared in order to stimulate rapprochement in phase 5. The Value Deliberation process is concluded in phase 6 by an evaluation in which the participants reflect on the process and how it influenced them. The five characteristics essential for legitimate deliberation of Fishkin (2009) apply to the Value Deliberation process: 1) during the preparation phase information and relevant data are distributed to all participants, 2) the steps to formulate the alternatives, arguments and conducting the value deliberation allow for comparing different positions and therefore provide substantive balance, 3) when inviting participants the initiator should ensure that the participants reflect all important perspectives so that diversity is reached, 4) an independent facilitator stresses the importance of conscientiously weighing all arguments, and 5) the facilitator should allow all participants to contribute to the discussion equally and underline that views are weighted on evidence and not on who proposes them. The Value Deliberation process meets Fishkin's five characteristics for legitimate deliberation and therefore we applied it for the value elicitation of the interpretation stage of the Glass Box Framework.

4.3 METHOD

Value elicitation in the context of Autonomous Weapon System deployment is qualitative research in which the participants interact and deliberate. The aim of the survey is to study if the value deliberation will change the participant's perception on the acceptability of the alternatives regarding a scenario of Autonomous Weapon System deployment. As the method for value elicitation, we chose the Value Deliberation process developed by Pigmans (2020), because it meets Fishkin's five characteristics for legitimate deliberation and it was tested in a large-scale citizen's summit event during the G1000 in July 2017 in Rotterdam (Pigmans, Dignum, & Doorn, 2021).

In our previous work on values related to Autonomous Weapon Systems, we studied people's perception on blame, trust, harm, human dignity, confidence, expectations, support, fairness and anxiety by comparing a scenario of the deployment of Human Operated drones to that of Autonomous Weapon Systems (Verdiesen, 2017; Verdiesen, Santoni de Sio, & Dignum, 2019). To select these values, we conducted a literature review, a short exploratory online survey and expert interviews. The values selected to incorporate in the Value Deliberation process in this research are based on our previous research as we find this the most complete overview of values related to Autonomous Weapon Systems.

4.4 RESEARCH SET-UP

Due to the COVID19 restrictions we designed an online value deliberation process instead of conducting the deliberation in person. We followed the process Pigmans (2020) described (Figure 13) and adjusted it to an online set-up consisting of a bipartite survey and a virtual session for the expert panel discussion. The first part of the survey was sent three days prior of the online discussion session and needed to be completed before the online session. The survey (see appendix A) started with the scenario and the options (the alternatives) that the Autonomous Weapon System could take were given (step 1 of Figure 14). Next, the participants were asked to list an advantage and disadvantage (the arguments) for each option, which is step 2, and rank the options from most acceptable to least acceptable (ranking 1 - step 3). During the online session the second part of the survey was sent to guide the value elicitation (step 4). For each option the participants were asked: *Which values are relevant for this option*? and *Are these values threatened or promoted in this option*? After filling in this part of the survey, the participants discussed values in the online session (step 5). Next, in step 6 the participants ranked the options a second time (ranking 2) in the survey. The online session concluded with a comparison and discussion on the ranking (step 7) and an evaluation (step 8). The advantage of the online setting is that the participants could join the survey from their own location which allowed for a diverse group with international participation without the need for travelling. The disadvantage of an online setting is that the non-verbal interaction and interpretation of facial expression is less clear than when conducting the session in person.

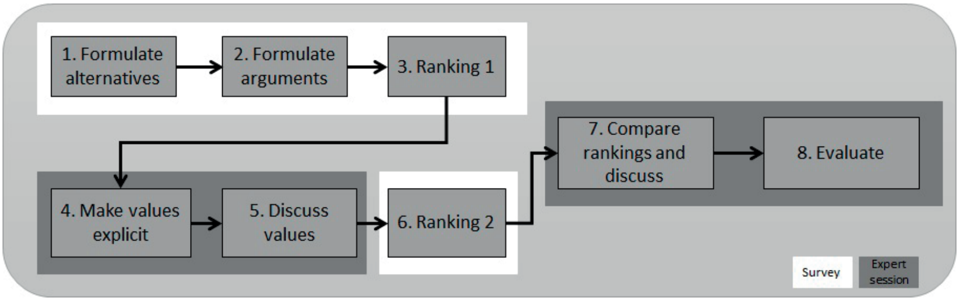


Figure 14: Value deliberation process differentiated in survey and expert online session (adapted from: Pigmans, 2020)

4.5 SCENARIO AND OPTIONS

The following scenario was used throughout the survey to describe the situation:

> *An Autonomous Weapon System provides force protection for soldiers that are clearing the road from improvised explosive devices. The Autonomous Weapon System is equipped with surveillance equipment, weapons (air-to-ground missiles) and flies autonomously in the Area of Operation. It is programmed to avoid flying over a restricted operating zone and an electronic warfare threat. The Autonomous Weapon System is equipped with facial and image recognition software for people, weapons and explosives. It is programmed with different options to engage when it recognizes a threat to the soldiers that are clearing the road. The Autonomous Weapon System detects movement behind a large rock near a narrow part of the road at a distance of 300 meters of the road clearance soldiers.*

After reading this scenario the participants read the options (the alternatives) each in turn and were asked to list an advantage and disadvantage per option (the arguments). These options were developed based on the military and technical domain knowledge of the primary researcher and discussed with a second researcher. During the pilot study these options were tested before using them in the actual study. The options presented to the participants are:

A. The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System warns the soldiers of the movement and takes no further action.

B. The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System asks permission to engage to neutralize the threat to the road clearance soldiers.

C. The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System automatically engages to neutralize the threat to the road clearance soldiers.
D. The Autonomous Weapon System positively identifies with a confidence of 99% using facial recognition all three persons sitting behind the rock as members of an opponent group aiming weapons at the road clearing soldiers. The Autonomous Weapon System automatically engages to neutralize the threat to the road clearance soldiers.
E. The Autonomous Weapon System positively identifies with a confidence of 99% using facial recognition all three persons sitting behind the rock as members of an opponent group aiming weapons at the road clearing soldiers. The Autonomous Weapon System shares the identification with the commander and asks permission to engage to neutralize the threat to the road clearance soldiers.
F. None of the options is acceptable.

4.6 SAMPLE PILOT AND ACTUAL STUDY

A pilot study was conducted before the actual survey and online session was held. The aim of the pilot study was to improve the research set-up and if possible, the results could be used in the survey. Eight researchers (PhD students and post-docs) participated in the pilot but due to a flaw in the set-up- the two questionnaires could not be linked - the results were not usable. However, the pilot study gave valuable insight in the usability of this set-up for the Value Deliberation process and allowed us to correct the problem for the actual study. The actual study was held in two separate sessions. We sent 33 invitations to experts on Autonomous Weapon System and 14 responded – a response rate of 42%. These experts were chosen based on their experience with, and knowledge of, autonomous systems. Most of them work or conduct research related to Autonomous Weapon System or in a closely related adjacent field. We divided the 14 in two groups to ensure that the group was not too large for people to contribute to the online value discussion. The participants were a mix of military personnel (21%) and civilians (79%) working at the Dutch Ministry of Defense (25%), an NGO (8%), researchers (33%), policymakers (17%) and industry (17%). Session 1 consisted of six participants and resulted in 5 usable results, because one participant had not filled in the questionnaire before online session. Session 2 consisted of 8 participants and resulted in 7 usable results. One participant finished questionnaire before value discussion and therefore the value discussion was not of influence on the ranking of the options which impacted the research results. The total number of usable results is n=12.We asked for some demographics; 93% of the participants has a university degree or PhD, 36% of the participants have worked with drones, 50% has worked with Artificial Intelligence and

36% has seen war or has been in a conflict zone.

The sample size (n=12) is not uncommon in qualitative studies. Studies have found extreme variations in sample size in qualitative research studies across all research designs (Marshall, Cardon, Poddar, & Fontenot, 2013). The sample size of a qualitative study can be determined by its *information power*. Information power depends on the aim of the study, sample specificity, use of established theory, quality of dialogue and analysis strategy. Information power indicates that the more information the sample holds, relevant for the actual study, the lower number of participants is needed (Malterud, Siersma, & Guassora, 2016). In our study, the panel consists of experts in the field of Autonomous Weapon System deployment. The aim of the study is narrow, the experts have high specific knowledge on the topic, the theoretical background is sufficient, the quality of the dialogue was strong and the analysis was done on a specific case (one scenario regarding the deployment of an Autonomous Weapon System). The information power of our sample is high and therefore the sample size is sufficient. We use the results to explore the effect of value deliberation on the acceptability of options for Autonomous Weapon System deployment to provide us with deeper insight into this real-world problem (Tenny, Brannan, Brannan, & Sharts-Hopko, 2022).

4.7 RESULTS

The nature of the data is qualitative so no statistical techniques are applied to analyse the results. The results are descriptive and are processed by using the Ranking-Calculator from the Value Deliberation Toolbox (https://www.delftdesignforvalues.nl/valuedeliberation-toolbox/). The data was processed after the online session so the participants could not reflect on it during the session. The results in Figure 15 show the ranking of the alternatives in round 1 (ranking 1 in Figure 15) and round 2 (ranking 2 in Figure 15). Ranking 1 is step 3 in the value deliberation process (Figure 14) and ranking 2 is step 6. The alternative with the lowest score is the most acceptable alternative and the alternative with the highest score is the least acceptable. The order from most to least acceptable alternatives in round 1 is: A, B, E, C, D, F. In round 2 the order is: A, B, E, D, C, F. Based on the value deliberation between ranking 1 and 2 a change in the order of the acceptability of alternatives is noticeable. The acceptability of the alternative C and D is flipped in round 2 compared to round 1. Although a minor change, it is interesting because the participants were asked at the end of the value deliberation if they changed their ranking order. Some participants indicated to have consciously changed the order, but most participants replied that they did not, or did not intended to, leaving the option open that the value discussion could have influenced their ordering. One participant mentioned that the value discussion changed the way she read the options. Based on the results it seems that some of the participants unconsciously changed the order of the acceptability of the alternatives.

Before the value deliberation, the participants were asked in part 2 of the survey for each of the alternatives: *which values are relevant for this option*? This is step 4- make values explicit- in the value deliberation process (Figure 14). They could check a predefined list of the values: fairness, suffering, accountability, responsibility, safety, harm, human dignity, meaningful human control, predictability, privacy, trust, reliability, proportionality, blame, robustness, explainability. These values were selected based on (Verdiesen, Santoni de Sio, & Dignum, 2019) and the pilot study in which the participants indicated which values they missed in the predefined list. The values that were highlighted as relevant for the alternatives were: safety, meaningful human control, proportionality, accountability, responsibility, predictability, reliability and explainability. As part of the evaluation (step 8 in Figure 14) participants were asked which values they missed on the predefined value list. Distinction, necessity, precaution, human autonomy, accuracy, human competences, relational and sociability between human and robot, mental and emotional health of the troops, usability and security were mentioned.

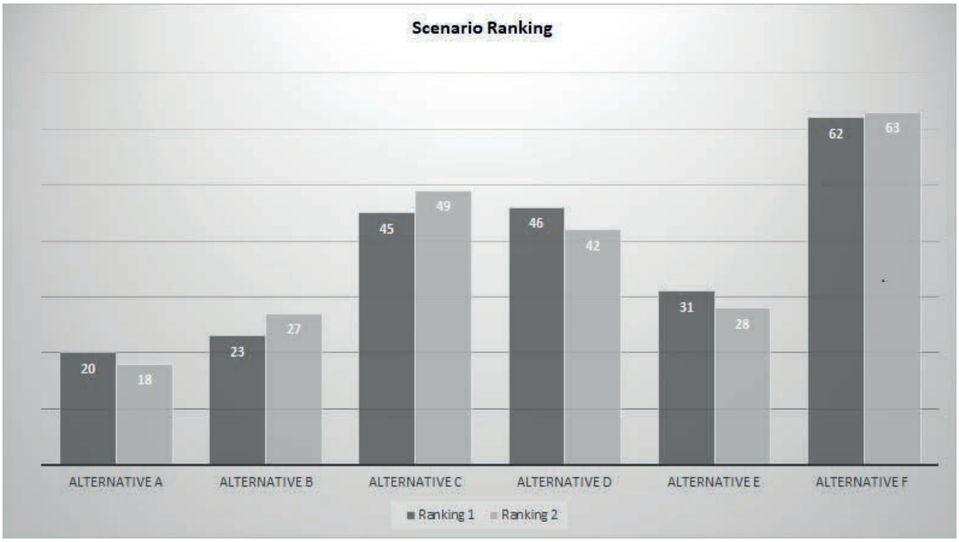


Figure 15: Overview results scenario ranking

At the start of part 1 of the survey and before the first ranking, the participants were asked to list an advantage and disadvantage of each alternative (step 2 in Figure 14). Early warning, safety of soldiers and quick response to threat were mentioned most as advantages. The disadvantages that were mentioned are: late response to threat, automation bias, false positives in identification and dehumanisation of the target. During the value deliberation (step 5 in Figure 14) experts from different backgrounds discussed the context of the scenario and alternatives. Their experience and background determined how they viewed the scenario and alternatives and influenced their answer and ranking. For example, a scientist viewed the values as being part of the design process, for a policymaker it was important that the system provides proper information and that the commander can review this information. One of the participants felt really uncomfortable with the image recognition and raised privacy issues in this 'big brother' scenario. An expert in computer vision viewed the 99% confidence as too uncertain, not reliable enough and not as an improvement of the system because it is more difficult to understand, but military personnel (nonexpert in computer vision) viewed the addition of 99% confidence as an increase in reliability of the system. Also, military personnel viewed the scenario based on the principles of the Rules Of Engagement and hostile intent which gave context to the scenario to base their answers on. This shows that the difference in experience and background, for example technical expertise or operational experience, influences the answers and ranking of the participants in the value discussion. This can impact design choices that are based on value elicitation so the variety of participant's background and level of expert knowledge should be taken into account when conducting the value deliberation and making design choices.

Another value that was discussed among the participants at the evaluation (step 8 in Figure 14) was *trust* in the system. One participant stated that compared to human decision-making an AI system can make decisions with fewer errors than human decision-making (for example with Autonomous Vehicles). The option in which the Autonomous Weapon System only was used as an early warning system was most acceptable and most trusted. Paraphrasing one of the military participants: 'It is about understanding the strategy and context of the mission. We need to understand the impact of technology and our presence on the mission. We should think better of applying which technology in which context.' This shows that not all applications of Autonomous Weapon System in a mission context provide trust to military experts in the decision-making of the Autonomous Weapon System. Human decision-making is in some cases more trusted and preferred. In general, the context in which an Autonomous Weapon System is deployed impacts the meaning and weight people attribute to the values associated with the Autonomous Weapon System.

4.8 VALIDATION RESULTS VALUE DELIBERATION PROCESS

We chose not to increase the sample size by holding additional value deliberation sessions to validate our results, because inviting laymen for this study will not provide additional qualitative data. In addition, we conducted an extra round of validation and invited four experts - who have not been part of the expert panel - to reflect on the results. Two experts responded and have reviewed the results and reflected on the usability for their field. Both experts indicated that the results are usable for their line of work and can apply the results in their work (see questionnaire in appendix B).

4.9 CONCLUSION

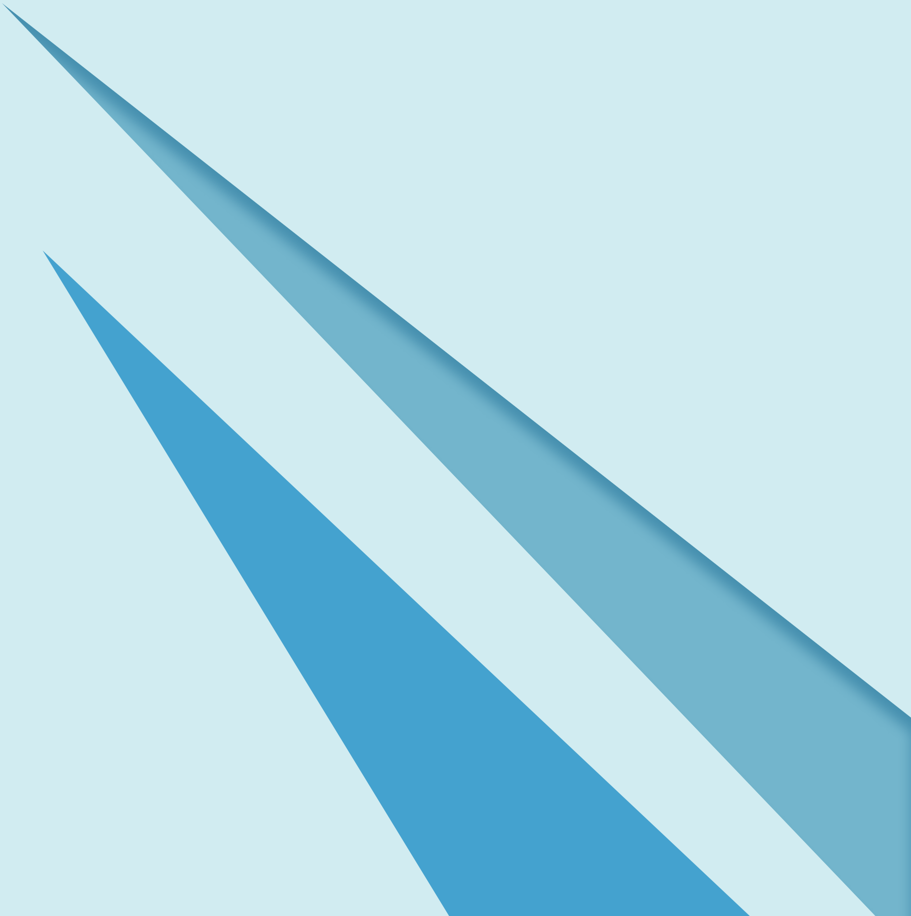
The value elicitation conducted using the Value Deliberation process not only shows that value discussion leads to changes in perception of the acceptability of alternatives in a scenario of Autonomous Weapon System deployment, it also gives insight into which values are deemed important and highlights that trust in the decision-making of an Autonomous Weapon System is crucial. As a next step in the interpretation stage of the Glass Box framework, norms and requirements can be derived based on this value elicitation. These requirements will feed into the observation stage as observable elements to monitor and verify. The review stage is required after deployment as an accountability process of which findings should feed back into the interpretation stage for a next deployment of an autonomous system and thereby close the loop between the stages.

The value discussion and evaluation disclosed that not all applications of Autonomous Weapon Systems in a mission context provide trust to military experts in the decision-making of the Autonomous Weapon System. Human decision-making is in some cases more trusted and preferred. In general, the context in which an Autonomous Weapon System is deployed impacts the meaning and weight people attribute to the values associated with the Autonomous Weapon System. The findings of this study imply that deliberate value discussion influences people perceptions of their values related to Autonomous Weapon Systems. More general, active participation in a value discussion leads to a conscious, and sometimes unconscious, change in people's preferences of alternatives. This could be beneficial in other areas than Autonomous Weapon Systems for policy making and citizen participation in local and national public administration. For example, to get citizen views on a municipal plan for the redevelopment of a local park or on a national level get input for nitrogen reduction policy. The application of the online Value Deliberation process method is not limited to Autonomous Weapon Systems and can be used in other areas as well.

Part IV

TECHNICAL INVESTIGATION PHASE

During the technical investigation phase, we operationalised the Glass Box framework by creating an implementation concept as an example to show that the Glass Box framework is actionable. We simulated the implementation concept using Coloured Petri Nets. The implementation concept is applied to the case of an autonomous military surveillance drone. We chose this application area, because not all researchers that were involved in this part of our research were comfortable with working on a scenario with an Autonomous Weapon System.



5|

Implementation concept

In this chapter we first introduce the scenario. Next, we describe Coloured Petri Nets: a discrete-event language for modelling synchronisation concurrency and communication processes that we used to simulate the implementation concept. We conclude with remarks on evaluating the implementation concept. Parts of this chapter have been published in Verdiesen et al. (2021).

5.1 SCENARIO

The scenario that we use in modelling the implementation concept is related to the scenario described in section 1.2. Before soldiers are sent out on a mission intelligence gathering with a drone, such as inspecting roads and clearing them of improvised explosion devices, is often conducted. In the scenario for designing the implementation concept, the autonomous drone is not weaponised and flies a surveillance mission over a deployment area to gather intelligence (see Figure 16). In addition, the drone should have a map to calculate its flight path. In this particular scenario it should remain within its Area of Operation and avoid certain areas, such as restricted operating zones and an electronic warfare threat.

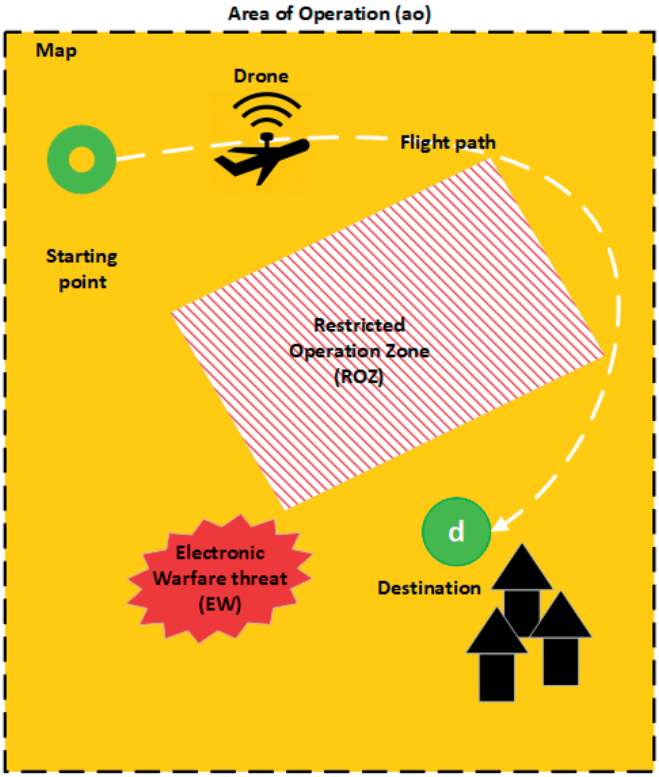


Figure 16: Visualisation scenario

In the first stage of the Glass Box framework (see Figure 10) the norms are derived from values before drafting (technical) requirements. Our implementation concept is based on existing operational norms within the Dutch Ministry of Defense, for example Rules Of Engagement, which always are available before the deployment of a mission. Therefore, value elicitation (block 1 of the Comprehensive Human Oversight Framework see Figure 7) is out-of-scope for our implementation concept. Value elicitation is described in chapter 4.

One of the norms (A) identified in the interpretations stage (block 4 in Figure 17) of our scenario is that the flight path should not cross a Restricted Operating Zone (ROZ). Another norm (B) is that the Electronic Warfare (EW) Threat should be avoided. The third norm (C) is that the surveillance drone should remain within the Area of Operation (AOO). These norms are input for the requirements (block 7 in Figure 17) for the drone's flight path, for example the drone should be able to plot waypoints based on GPS.

The requirements are translated to inputs for the monitoring in the observation stage (block 2 in Figure 17), e.g. GPS coordinates of the ROZ, EW Threat and AOO. After the mission, the norms are verified by manually evaluating the flight path (block 8 in Figure 17) to check if the autonomous drone has stayed within the AOO, did not cross the ROZ and EW Threat. Violation of the norms is reported in the mission debrief report as part of the review stage of the accountability process in which the conduct of the system (flightpath of the drone- block 9 in Figure 17), the actor (the drone- block 6 in Figure 17) should be discussed in a forum (mission debrief- block 3 in Figure 17).

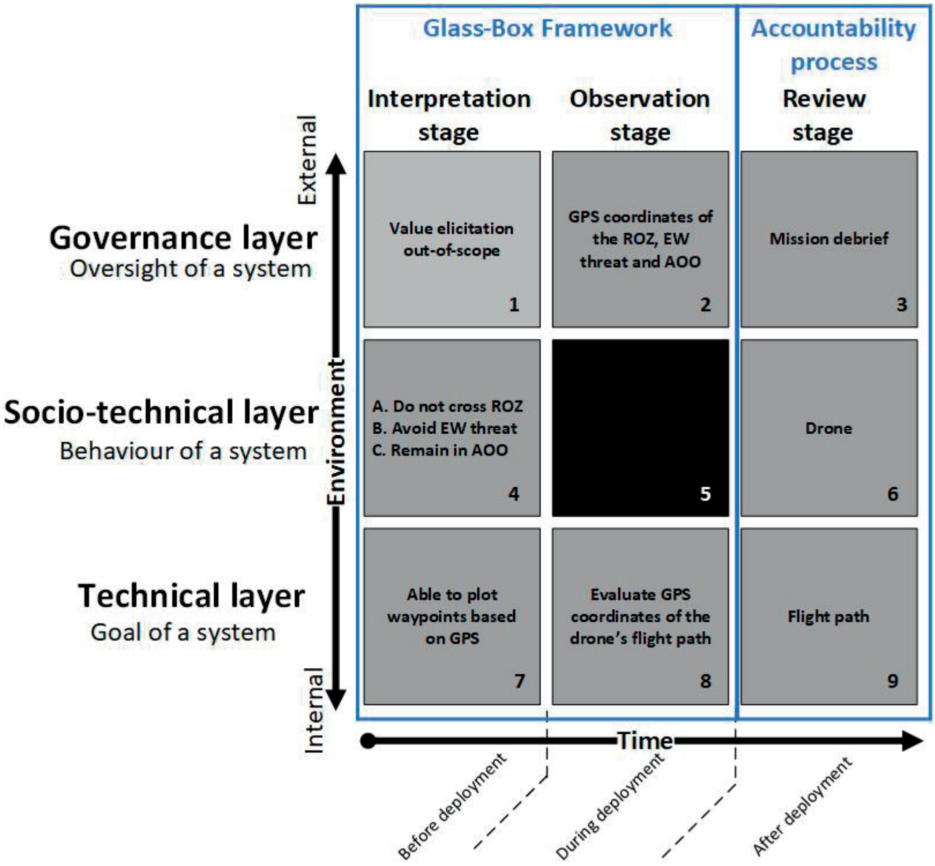


Figure 17: Scenario implementation concept plotted on the Glass Box framework and the Comprehensive Human Oversight Framework

5.2 SIMULATION OF IMPLEMENTATION CONCEPT

We created a simulation of the implementation concept of a pre- and post-flight procedure as an example using Coloured Petri Nets (CPNs) as modelling language (Jensen, 1994). CPNs is a discrete-event language for modelling synchronisation concurrency and communication processes. The language consists of states and events and a system that can change a state. CPNs combine a description of the synchronisation of concurrent processes with the primitives of a programming language which enables the definition of data types and adjustment of data values. CPNs have been applied to a variety of systems, from the description of work processes, communication protocols, distributed algorithms to flexible manufacturing processes (Jensen, 1994). Jensen (1994) provides two reasons for applying CPNs. The first is that a CPN model can specify or present a system which allows us to explain it to ourself or other people and investigate it before it is constructed. Secondly, it dramatically increases the understanding of the modelled system which is often more beneficial than the description or analysis results themselves. Both reasons are our motives to model the implementation concept using CPNs.

We used CPN Tools to create a simulation that allows us to check the monitoring and verification process of the observation stage of the Glass Box framework and run a simulation-based performance analysis (Jensen, Kristensen, & Wells, 2007). CPN Tools has several settings to conduct an automatic simulation based on a random number generator to calculate the effects of occurring steps (Jensen, 1994). We created a simulation that shows the steps of a pre-flight mission planning and post-flight mission evaluation process for autonomous surveillance drones which is not too complex as an example. We based the processes on the scenario described in section 1.2. As reference we used information obtained in several conversations with domain experts in the Dutch Ministry of Defense and the JFCOM-UASPocketGuide-the US Army Unmanned Aerial Systems manual (JUAS-COE, 2010). The CPNs are uploaded as Supplementary Materials (https://github.com/responsible-ai/DroneCPN).

Pre-Flight Mission Planning Process

In the pre-flight mission planning process, first the steps are modelled to check the prerequisites for a mission; i.e., the availability of a map and the status of the weather conditions (see Figure 18). Next the compliance criteria Area of Operation, Restricted Operating Zone, and Electronic Warfare Threat are checked and if these are complied with, the flight path is calculated. If, for example, the boundaries of the Area of Operation are not known and this criterion is not complied with, then the process enters a feedback loop in which the boundaries of the Area of Operation are requested. When all criteria are met the approval process is triggered and sequentially a drone is requested. In the case that the mission is not approved, the reason for disapproval needs to be solved first in order to continue the process. The pre-flight mission planning process is modelled with several feedback loops. For example, if there is no map available then a map is requested or if the weather conditions are adverse than the mission is replanned (Figure 18). In the final step the mission is flown and, upon completion of all the steps, the pre-flight mission planning process ends and the drone is returned to the pool of drones and can be deployed for a next mission (see Figure 19 and Figure 20).

Post-Flight Mission Evaluation Process

The evaluation of the mission will be done manually and starts with two concurrent steps. The check of (1) the compliance criteria and (2) the flight path. The same compliance criteria as in the pre-flight mission planning process are checked (see Figure 21); Area of Operation, Restricted Operating Zone and Electronic Warfare Threat. If the criteria, for example, “avoid Restricted Operating Zone”, is met, the process passes to the next stage. If the Restricted Operating Zone is crossed, the criterion is not met and this norm violation will be noted in the debrief report. Concurrent to this step, the compliance with the flight path, or deviation of it, will be checked. Both compliance with the criteria and the flight path as noncompliance will end up in the debrief report. Noncompliance comments can be used as lessons learned for the next mission. The draft of the debrief report is the final step of the post-flight evaluation process (see Figure 22) and this evaluation can be used in the review stage of the accountability process.

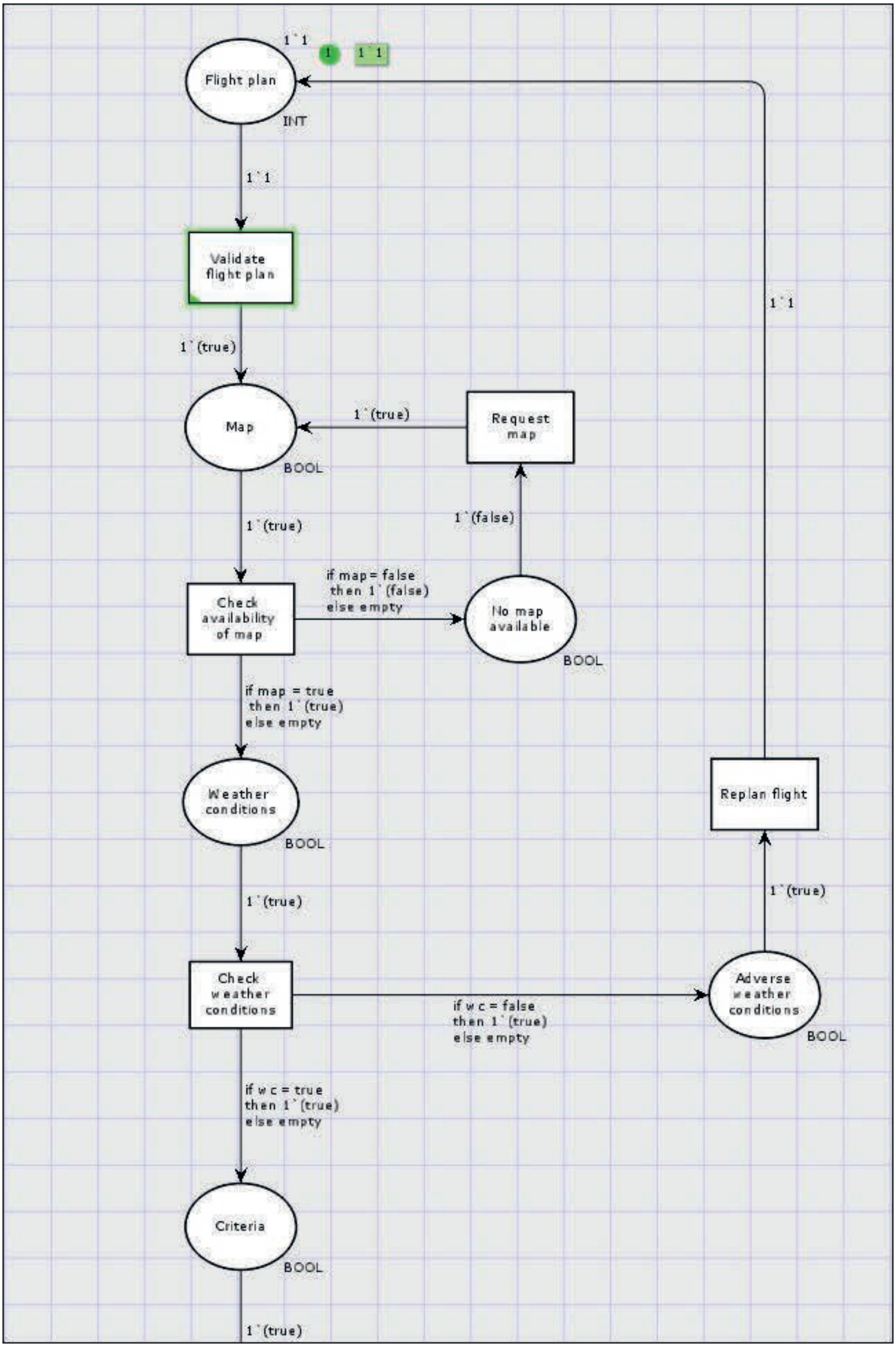


Figure 18: Pre-Flight Mission Planning Process: screenshot 1

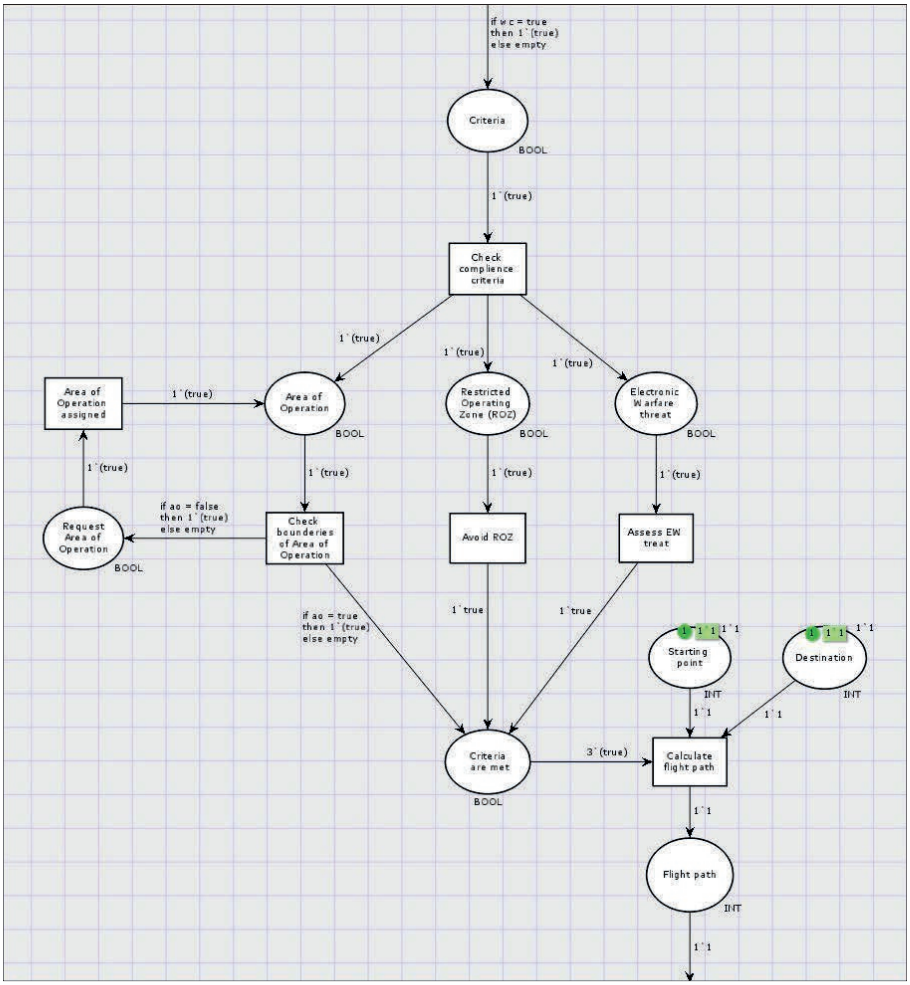


Figure 19: Pre-Flight Mission Planning Process: screenshot 2

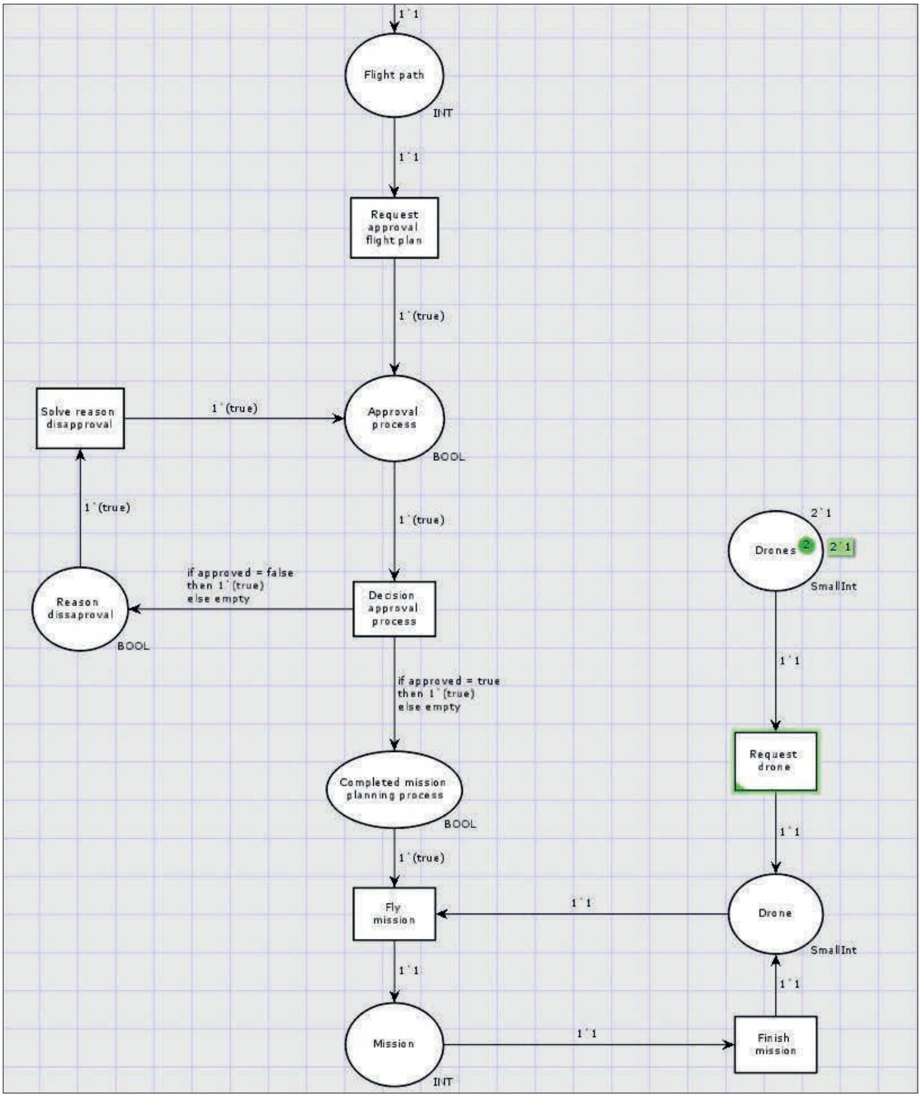


Figure 20: Pre-Flight Mission Planning Process: screenshot 3

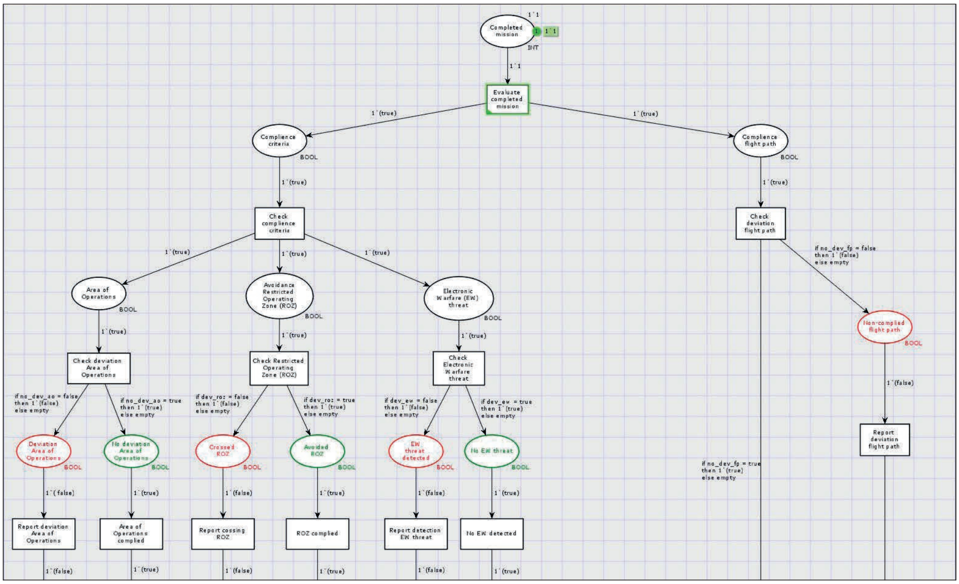


Figure 21: Post-Flight Mission Evaluation Process: screenshot 1

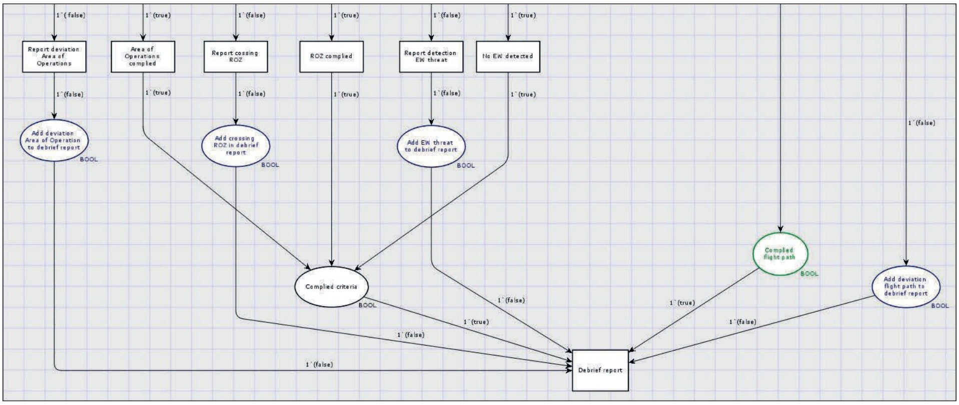


Figure 22: Post-Flight Mission Evaluation Process: screenshot 2

5.3 EVALUATION OF SIMULATION OF THE IMPLEMENTATION CONCEPT

The CPNs in figures 18 to 22 have been used to simulate the monitoring and verification process of the observation stage of the Glass Box framework (which is described in section 3.4). In general, a simulation imitates a process by creating another process of which an object or system changes its state in time. Simulations can be used for different reasons: '*as a technique to investigate the detailed dynamics of a system, as a heuristic tool to develop hypotheses, models and theories, as a substitute for an experiment to perform numerical experiments, as a tool for experimentalists to support experiments and finally as a pedagogical tool to gain an understanding of the process*' (Hartmann, 1996, p. 6).

In our case, we have created the CPNs as a means to gain an understanding of the observation stage of the Glass Box framework by visualizing the processes and to check if it is possible to apply the monitoring and verification process to a practical case of autonomous surveillance drones by modelling the Pre-Flight Mission Planning Process and Post-Flight Mission Evaluation Process. Evaluating this simulation by verification of the behavioural properties of CPN models can be conducted using a state space exploration. A state space is a directed graph consisting of a node for each marking and edges corresponding to the events. By computing all reachable states and state changes, it is possible to verify questions regarding the behaviour of a system. The modeler can add tokens to the CPN model to ensure a finite state space so that the behavioural properties can be checked (Jensen, Kristensen, & Wells, 2007).

For the evaluation of our simulation, the ASCoVeCo State Space Analysis Platform (ASAP) - developed by Westergaard, Evangelista, and Kristensen (2009)- could be used. ASAP is a tool that supports state space exploration and analysis of CPNs. The ASAP architecture consists of a Graphical User Interface (GUI) and State Space Exploration Engine (SSE engine) that is implemented in Java based on the Eclipse platform. The GUI allows creating and managing *verification projects* consisting of *verification jobs*. ASAP could be used to evaluate and verify the behavioural properties of our CPNs. However, to conduct the evaluation with the ASAP architecture the user will need to be experienced with Java and Eclipse to upload the CPNs to the SSE engine.

Bearing in mind that not all researchers are proficient with Java and Eclipse, a second approach for evaluation of the simulation could be by querying experts in military drone deployments to verify if the simulated processes represent an actual Pre-Flight Mission Planning Process and Post-Flight Mission Evaluation Process. The evaluation can be conducted by either organizing expert panels (for example based on the Delphi method to structure the process) or one-on-one sessions with experts. This qualitative evaluation should preferably be conducted by independent researcher that was not involved in this research for a critical review of the model.

CPNs have been used for over forty years to model different processes in different application domains such as manufacturing, computer networks and even a NORAD[1] command post (Shapiro, Pinci, & Mameli, 1993). Our simulation is applied to a very specific use case - that of military surveillance drones - and it is not clear if it will also be applicable to other autonomous systems as well. This is one of the limitations of our research (see section 7.2). In further research the CPNs should be extended to other scenarios of autonomous systems in the military domain in order to check if the CPN approach is scalable and can be extended to a broadly used tool for oversight of autonomous systems.

5.4 CONCLUSION

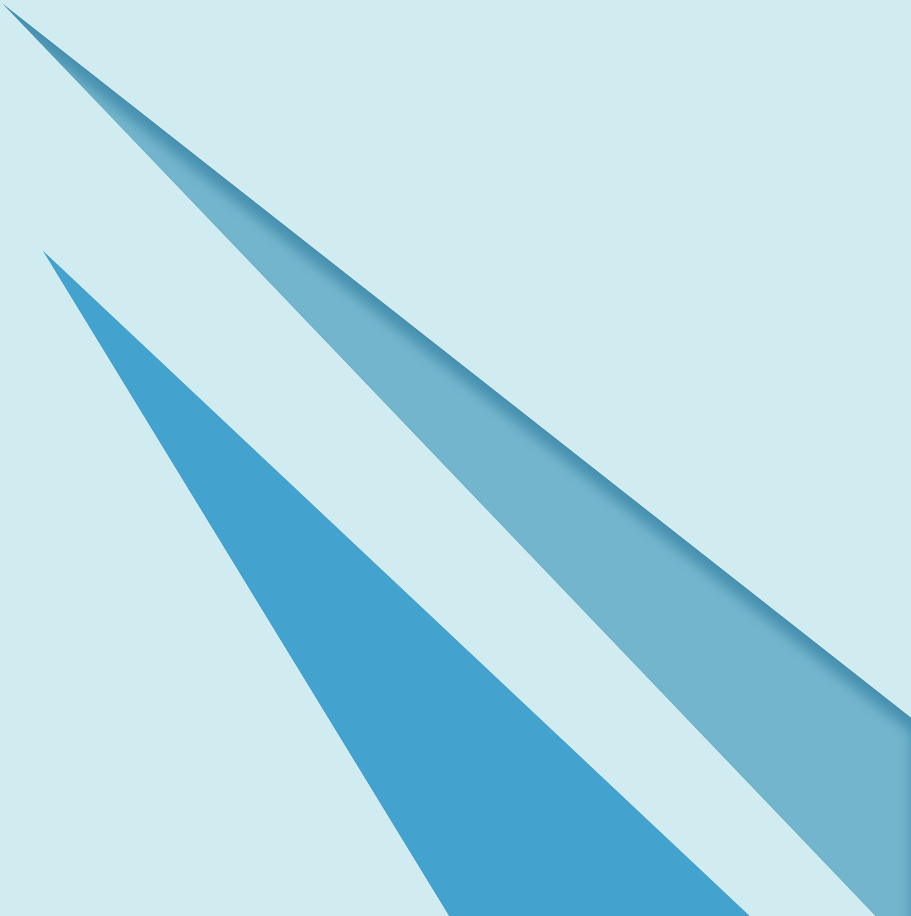
The simulation of the implementation concept shows that it is possible to set criteria in the pre-flight process and to evaluate these criteria post-flight. During flight, the drone itself is treated as a black box of which the internal logic is not accessible. Although being a toy example, it demonstrates that a monitoring process can be designed to implement human oversight where the users set norms - or criteria in this example - for input and observe and evaluate the output against the input to check for noncompliance of the norms. Deviations of the norms will be reported in the verification process and can be used to update the norms in a new scenario. This way the users do not need technical skills to understand the internal workings of drone, but still can monitor and oversee the use of the autonomous system based on observable norms. We do not monitor the in-flight actions of the autonomous drone in our implementation concept, because we assume that in-flight communication is not possible, for example due to a failing communication structure, an Electronic Warfare Threat, or operator unpreparedness. Therefore, it is not possible to oversee norm violations nor is it possible to intervene during the flight as would often be the case with black box autonomous systems.

[1] NORAD: North American Aerospace Defense

Part V

CONCLUSION AND DISCUSSION

The objective of this research is to improve the allocation of accountability and responsibility in the deployment of Autonomous Weapon Systems by designing a framework and implementation concept such that the criteria for Human Oversight are identified, represented and validated. In chapter 6, the results of this research are described answering the research questions presented in chapter 1. In chapter 7 we highlight the emerged insights over the past five years, followed by the limitations of this research and suggestions for future work. We conclude this chapter by presenting the contributions and summarizing the policy recommendations that we made throughout this dissertation.



6|

Conclusion

In this chapter we follow the three phases of our research approach to answer our research questions (section 1.1) based on the results of our research.

6.1 CONCEPTUAL INVESTIGATION PHASE

In the conceptual investigation phase first the delineation of the values accountability and responsibility and a theoretical view on the concept of Human Oversight for Autonomous Weapon Systems are described. Next, a framework for Comprehensive Human Oversight is designed to give an overview of the control mechanisms and show the gaps that emerge when introducing Autonomous Weapon Systems.

The answer to the first part of the question:

Q1 *What are Autonomous Weapon Systems…?*

is that Autonomous Weapons Systems are *'A weapon that, without human intervention, selects and engages targets matching certain predefined criteria, following a human decision to deploy the weapon on the understanding that an attack, once launched, cannot be stopped by human intervention.'*(AIV & CAVV, 2015, p. 11; Broeks et al., 2021, p. 11). In our opinion this definition of the ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS (AIV & CAVV) captures the description of Autonomous Weapon Systems best from an engineering and military standpoint, because it takes predefined criteria into account and is linked to the military targeting process as the weapon will only be deployed after a human decision.

To answer the second part of the question

… and how are the values of accountability and responsibility related to the concerns for the deployment of Autonomous Weapon Systems?

we have identified several concerns that are mentioned in the societal and academic debate on Autonomous Weapon Systems. Next to security risks and unpredictable activities, the impact on human dignity and the emergence of an accountability gap are mentioned as concerns with the use of Autonomous Weapon Systems. The alleged offence to human dignity entailed in delegating life-or-death decision-making to a machine is linked to the value of human life. Also, many scholars express concerns that an accountability gap or accountability vacuum will emerge when Autonomous Weapon Systems are deployed. An accountability gap or vacuum arises when no human can be held accountable for the decisions, actions and effects of Autonomous Weapon Systems (Matthias 2004; Asaro 2012; Asaro 2016; Crootof 2015; Dickinson 2018; Horowitz and Scharre 2015; Wagner 2014; Sparrow 2016; Roff 2013; Galliott 2015).

The concerns described above highlight that responsibility and accountability are values often mentioned in the societal and academic debate around Autonomous Weapon Systems. Responsibility can be forward-looking to actions to come and backward-looking to actions that have occurred. Accountability is a form of backward-looking responsibility, refers to the ability and willingness of actors to provide information and explanations about their actions and defines mechanisms for corporate and public governance to hold agents and organisations accountable in a forum. Responsibility contributes to minimizing unintended consequences by anticipating on actions and unintended consequences to come and taking measures to prevent or mitigate them. Accountability can decrease unintended consequences in providing information and explanations by actors of their previous actions in order for other actors to learn from them and prevent mistakes and unintended consequences of their own.

To answer research question two:

Q2 *How should the values of accountability, responsibility and the concept of Human Oversight be characterized?*

we turned to philosophical, political science, public management, international relations, social psychology, constitutional law and business administration literature. By reviewing the literature of these fields, we found that the term accountability has two different uses. On the one hand, it is used in a broad sense to praise or criticize the performance of states, organizations, firms or officials regarding policy or decisions in relation to their ability and willingness to give information and explanations about their actions ('accountability as a virtue'). On the other hand, in a narrow sense, accountability is also used to define the mechanisms for corporate and public governance to hold agents and organisations accountable ('accountability as a mechanism') (Bovens, Schillemans, & Goodin, 2014). Accountability is not only scrutiny after the event has occurred, it also has a preventive and anticipatory use for which norms are (re)produced, internalized and adjusted by means of accountability if necessary.

Responsibility can be forward-looking to actions to come and backward-looking to actions that have occurred. Van de Poel (2011) focusses on moral responsibility for consequences to describe the notions of forward- and backward-looking responsibility and does not describe organizational, social and legal responsibility nor responsibility for actions. Two varieties of responsibility that are primarily forward-looking are: 1) responsibility as virtue and 2) the moral obligation that something is the case; and three varieties that are primarily backward-looking are: 3) accountability, 4) blameworthiness and 5) liability. Following this reasoning we found that accountability is backward-looking and part of the responsibility which encompasses more than accountability alone.

The concept of Human Oversight has been researched by several scholars who mention that an oversight mechanism is needed in order to hold an actor accountable (Caparini, 2004; Schedler, 1999; Scott, 2000). West and Cooper (1989: in (Pelizzo, Stapenhurst, & Olson, 2006)) mention two reasons for oversight in the political system: (1) it can improve the quality of policies or programs and (2) when policies are ratified by the legislative branch, they obtain more legitimacy. The oversight mechanism can be implemented as an *ex post* review process or a mechanism for either *ex post* of *ex ante* supervision (Pelizzo et al., 2006). Oversight over international institutions can be used as an equivalent for the accountability of these institutions according to De Wet (2008).

As Bovens (2007) notes, accountability can be viewed as a form of control, but not all forms of control are accountability mechanisms. Therefore, we turned to our third research question:

***Q3** Which control mechanisms are described in literature and present in the military domain, and which gaps in control mechanisms can be identified by the introduction of Autonomous Weapon Systems?*

Control has traditionally been defined in different ways, depending on application domain. Control from an engineering perspective can be described as a mechanism that compares the output of another system or device to the input and goal function by means of a feedback loop to take action to minimize the difference between outcome and goal. The traditional engineering perspective holds a very mechanical or cybernetic view on the notion of control, one that is not well-suited to make sense of the interaction between a human agent and an intelligent system for which the human is to remain accountable.

The socio-technical perspective on control describes which agent has the power to influence the behaviour of another agent (Koppell, 2005). An agent can be human or a technological system. The influence of one agent over another is often mediated by technology and it also includes controlling the technology. Scott (2000) makes a distinction between *ex ante* and *ex post* control. *Ex ante* involvement in decision-making is related to managerial control and accountability-based control is linked to *ex post* oversight. Control from a socio-technical perspective is power-oriented and aimed to influence behaviour of agents making use of *ex ante*, *ongoing* or *ex post* instruments. However, it does not explicitly include mechanisms of power over nonhuman intelligent systems, like Autonomous Weapon Systems.

The governance perspective on control describes which institutions or forums supervise the behaviour of agents to govern their activities. Pesch (2015) argues that there is no institutional structure for engineers which calls on them to recognize, reflect upon and actively integrate values into the designs on a structural basis. The result is that the moral effects of a design can only be evaluated and adjusted after the implementation in society. Pesch (2015) notes that engineers relate to different institutional domains, such as the market, the state and science. The consequence is that engineers do not have a clearly defined accountability forum and that they rely on engineering ethics and codes of conduct. However, these codes of conduct are often not robustly enough institutionalized to be regarded as a good regulative framework. Therefore, engineers use methods such as the Value-Sensitive Design and Constructive Technology Assessment as proxies for accountability forums. The need to develop and use these proxies for engineering practices reveals that a governance perspective on responsibility and control lacks robust institutionalized frameworks.

From a military perspective, control is described as a process to check if current and planned orders are on track and if the objectives to achieve a goal are met (Alberts & Hayes, 2006; Liao, 2008; NATO, 2017). Control aims to make adjustments to the plan if the current state deviates from the planned end-state of the mission. Control measures bound the mission space by limiting the area of operation, duration of military operations and by defining the order of battle. Control consists of procedures for planning, directing and coordination of resources for a mission and this includes standard operating procedures (SOPs), Rules Of Engagement (ROEs), regulations, military law, organizational structures and policies (Pigeau & McCann, 2002). Control in a military perspective is an instrument to bound and check if the actions are in line with the planned military goal and to adjust the planning when the current state deviates from the end state. This resembles the notion of control in an engineering perspective because there is a goal, input and feedback loop to adjust the system.

The insufficiency of traditional notions of control to make sense of the human control over Autonomous Weapon Systems required to ground accountability, has led to the introduction of the notion of Meaningful Human Control in the political debate on Autonomous Weapon Systems. However, a common definition of this notion has been lacking in practice for a long time (Ekelhof, 2019). Often the notion of Meaningful Human Control has a very operational view and is strongly, if not exclusively, focused on the relation between one human controller and one technical system, and tries to identify the different conditions under which that controller may be able to effectively interact with the system. We may call this a narrow notion of Meaningful Human Control, insofar as the broader perspective of governance of control, organisational aspects, values and norms does not seem to be incorporated.

In an attempt to overcome the conceptual impasse on the notion of Meaningful Human Control, Santoni de Sio and Van den Hoven (2018) tried to offer a deeper philosophical analysis of the concept, by connecting it more directly to some coming from the philosophical debate on free will and moral responsibility. Mecacci and Santoni De Sio (2019) operationalized this concept of Meaningful Human Control even further in order to specify design requirements. Their framework shows that the narrow focus of engineering and human factors control needs to be widened to allow a development of autonomous technologies that are sufficiently responsive to ethical and societal needs. We may call this broad Meaningful Human Control. In recent years, several scholars have been working on operationalising the concept of Meaningful Human Control (Amoroso and Tamburrini, 2021; Umbrello, 2021; Cavalcante Siebert et al., 2022). All of these approaches have in common that they focus on the human-machine interaction in order to operationalise the concept of Meaningful Human Control. Either by bridging the gap between weapon usage and ethical principles based on 'if-then' rules (Amoroso & Tamburrini, 2021), creating actional properties for the design of AI systems in which each of the properties human and artificial agents interact (Cavalcante Siebert et al., 2022), or proposing two LoA's in which different agents have different levels of control over the decision-making process to deploy an Autonomous Weapon System (Umbrello, 2021). However, the wider conception of the control loop mentioned above does not incorporate the social institutional and design dimension at a governance level. The governance level 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. Therefore, we look at an oversight mechanism to connect the technical, socio-technical and governance perspective of control which may ensure solid controllability and accountability for the behaviour of Autonomous Weapon Systems. To connect these perspectives, we propose a Framework for Comprehensive Human Oversight that broadens the view on the control over Autonomous Weapon Systems and take a comprehensive approach that goes beyond the notions of control described above.

6.2 EMPIRICAL INVESTIGATION PHASE

In the empirical investigation phase, we answered research question four:

> ***Q4*** *To what extent can an empirical study be used to elicit values and how does this discussion lead to changes in perception of values of accountability and responsibility in a scenario of Autonomous Weapon System deployment?*

We used the Value Deliberation Process developed by (Pigmans, 2020) as an empirical study to identify values of experts. The experts could select values on a predefined list in the survey that was part of the Value Deliberation Process. The values that were highlighted as relevant were: safety, meaningful human control, proportionality, accountability, responsibility, predictability, reliability and explainability. As part of the evaluation the participants were asked which values they missed on the predefined value list. Distinction, necessity, precaution, human autonomy, accuracy, human competences, relational and sociability between human and robot, mental and emotional health of the troops, usability and security were mentioned.

The value discussion and evaluation disclosed that not all applications of Autonomous Weapon Systems in a mission context provide trust to military experts in the decision-making of the Autonomous Weapon System. Human decision-making is in some cases more trusted and preferred. The value elicitation conducted using the Value Deliberation Process not only gives insight into which values are deemed important and highlights that trust in the decision-making of an Autonomous Weapon System is crucial, it also shows that value discussion leads to changes in perception of the acceptability of alternatives in a scenario of Autonomous Weapon System deployment. The context in which an Autonomous Weapon System is deployed impacts the meaning and weight people attribute to the values associated with the Autonomous Weapon System. The findings of this research imply that deliberate value discussion influences people perceptions of their values related to Autonomous Weapon Systems. More general, active participation in a value discussion leads to a conscious, and sometimes unconscious, change in people's preferences of alternatives. Therefore, we can conclude that a value discussion can identify values and leads to changes in perception of values related to Autonomous Weapon Systems.

6.3 TECHNICAL INVESTIGATION PHASE

In the technical investigation phase, we took the next step and created observable criteria based on the Glass Box framework to answer question five:

> ***Q5*** *To what extent can Human Oversight be translated into observable criteria for the deployment of Autonomous Weapon Systems?*

The Glass Box approach (Aler Tubella, Theodorou, Dignum, & Dignum, 2019) is a framework for monitoring adherence to the contextual interpretations of abstract values which focuses uniquely on the observable inputs and outputs of an intelligent system. Its focus on the observable aspects of the system's behaviour makes it particularly apt for monitoring autonomous and generally opaque systems.

The Glass Box approach consists of two phases which inform each other: interpretation and observation. The interpretation stage consists of a progressive process of concretising abstract values into specific design requirements. The output from the interpretation stage is an abstract-to-concrete hierarchy of norms where the highest level is made up of values and the lowest level is composed of fine-grained concrete requirements for the intelligent system only related to its inputs and outputs. The second phase of the approach is given by the observation stage. This stage is informed by the requirements on inputs and outputs identified in the interpretation stage, as they determine what must be verified and checked. In the observation stage, the system is evaluated by studying its compliance with the requirements identified in the previous stage. Feedback between interpretation and observation stage throughout the lifespan of the system is necessary: continuous observation informs us on which requirements are consistently unfulfilled, which may prompt changes in the implementation or in the chosen requirements.

We projected the two stages of the Glass Box framework on the Comprehensive Human Oversight Framework. The Interpretation stage of the Glass Box framework, in which values in the governance layer are turned into concrete norms, constraining observable elements and actions in the socio-technical layer, which in turn are translated into requirements in the technical layer, is done before deployment. During deployment the behaviour and actions of an autonomous system are monitored in the governance layer and verified in the technical layer in the Observation stage of the Glass Box framework that treats the block in the socio-technical layer as a black box. After deployment a Review stage is required as an accountability process in which a forum in the governance layer can hold an actor in the socio-technical layer accountable for its conduct in the technical layer. The outcome of the Review stage should feed back into the Interpretation stage for a next deployment of an autonomous system and thereby close the loop between the stages.

In this research we based our implementation concept on existing operational norms within the Dutch Ministry of Defense, for example Rules of Engagement, the limits of an Area of Operation and Restricted Operating Zone, which are available before the deployment of a military mission. Although we conducted a value elicitation in our empirical investigation phase, we did not translate these values into the observable norms of our implementation concept. However, we are convinced that is theoretically and practically possible to translate these identified values in observable norms, but the extent of this needs to be researched.

Our sixth and final research question:

> ***Q6*** *To what extent can observable criteria for Human Oversight be incorporated in an implementation concept for the deployment of Autonomous Weapon Systems?*

was answered by designing and building an implementation concept to operationalise the Glass Box framework as an example to show that the framework is actionable. We have created he CPNs as a means to gain an understanding of the observation stage of the Glass Box framework by visualizing the processes and to check if it is possible to apply the monitoring and verification process to a practical case of autonomous surveillance drones by modelling the Pre-Flight Mission Planning Process and Post-Flight Mission Evaluation Process.

The implementation concept shows that it is possible to set observable criteria in the pre-flight process and to evaluate these criteria post-flight. During flight, the drone itself is treated as a black box of which the internal logic is not accessible. Although being a toy example, it demonstrates that a monitoring process can be designed to guarantee human oversight where the users set norms - or criteria in this example - for input and observe and evaluate the output against the input to check for noncompliance of the norms. Deviations of the norms will be reported in the verification process and can be used to update the norms in a new scenario. This way the users do not need technical skills to understand the internal workings of a drone, but still can monitor and oversee the use of the autonomous system based on observable norms. We do not monitor the in-flight actions of the autonomous drone in our implementation concept, because we assume that in-flight communication is not possible, for example due to a failing communication structure, an Electronic Warfare Threat, or operator unpreparedness. Therefore, it is not possible to oversee norm violations nor is it possible to intervene during the flight as would often be the case with black box autonomous systems. Based on this part of our research we can conclude that it is possible to simulate observable criteria in an implementation concept for the deployment of Autonomous Weapon Systems.

7|

Discussion

In this chapter we highlight the emerged insights over the past five years, followed by the limitations of this research and suggestions for future work. We conclude this chapter by presenting the contributions and summarizing the policy recommendations that we made throughout this dissertation.

7.1 EMERGING INSIGHTS

Over the course of this research spanning the past five years, other scholars have published their recent research and new insights have emerged. During the update of our literature review we discovered two developments touching on our research that are worth mentioning. First a proposal for a value-neutral definition by Taddeo and Blanchard (2022) and second the work on defining the concept of Meaningful Human Control continued. We will provide a brief overview of both developments and briefly analyse the implications on our research.

Value-neutral definition of Autonomous Weapon Systems
In their comparative analysis of twelve existing definitions of Autonomous Weapon Systems by states or key international actors, Taddeo and Blanchard (2022) found that these definitions emphasize different aspects of Autonomous Weapon Systems and therefore lead to different approaches to address legal and ethical challenges with these type of weapon systems. They provide a value-neutral definition of Autonomous Weapon Systems based on four aspects: 1) autonomy, 2) adapting capabilities, 3) human control, 4) purpose of use – all of which are key according to them when considering ethical and legal challenges. The definition they drafted reads as follows:

> ‘*an artificial agent which, at the very minimum, is able to change its own internal states to achieve a given goal, or set of goals, within its dynamic operating environment and without the direct intervention of another agent and may also be endowed with some abilities for changing its own transition rules without the intervention of another agent, and which is deployed with the purpose of exerting kinetic force against a physical entity (whether an object or a human being) and to this end is able to identify,* ***select or attack the target without the intervention of another agent*** *is an AWS. Once deployed, AWS can be operated with or without some forms of human control (in, on or out the loop). A lethal AWS is specific subset of an AWS with the goal of exerting kinetic force against human beings.*’ (Taddeo & Blanchard, 2022, p. 15).

We agree with the conclusion of Taddeo and Blanchard (2022, p. 18): ‘*The debate on AWS is shaped by strategic, political, and ethical considerations. Competing interests and values contribute to polarize the debate, while politically loaded definitions of AWS undermine efforts to identify legitimate uses and to define relevant regulations.*’ The value-neutral definition of Autonomous Weapon Systems that they offer is a valuable addition to the academic and political debate. The implication for our research is limited, because only one aspect (highlighted in both definitions) is similar as the definition of the AIV&CAVV that we adhere to in this research:

> *'A weapon that, without human intervention,* ***selects and engages targets*** *matching certain predefined criteria, following a human decision to deploy the weapon on the understanding that an attack, once launched, cannot be stopped by human intervention.'*(AIV & CAVV, 2015, p. 11; Broeks et al., 2021, p. 11).

The definition of the AIV & CAVV explicitly mentions that targets should match predefined criteria and that the weapon will be deployed following a human decision. These two aspects are missing in the value-neutral definition presented by Taddeo and Blanchard (2022). However, from a military perspective these two aspects are imperative for responsible use of Autonomous Weapon Systems. Although the definition of Taddeo and Blanchard (2022) is a valuable addition to the academic and political debate on Autonomous Weapon Systems, we still adhere in our research to the definition of the AIV &CAVV for the reasons mentioned above.

Operationalising Meaningful Human Control

In recent years, several scholars have been working on operationalising the concept of Meaningful Human Control. Amoroso and Tamburrini (2021) created a normative framework for Meaningful Human Control. They suggest a differentiated approach and to abandon the search for a one-size-fits-all solution. Three roles are described by them in order for human control over weapon systems to be meaningful: 1) '…*human operators must play the role of fail-safe actor, preventing malfunctioning weapons from resulting in direct attacks against civilian populations or excessive collateral damages*', 2) *'…human control must function as an accountability attractor, securing legal conditions for criminal responsibility ascription in case a weapon follows a course of action that is in breach of international law.*' and 3) *'…human control operates as a moral agency enactor, ensuring that decisions affecting the life, physical integrity, and property of people involved in armed conflicts, including combatants, are not taken by artificial agents.'*(Amoroso & Tamburrini, 2021, p. 258). They state that rules are needed to bridge the gap between specific weapon systems and their uses on one hand and the ethical and legal principles on the other hand. These rules could be represented as "if-then" statements: the 'if' statement includes properties regarding the *what* and *where* of the mission and *how* it will perform its task, and the 'then' statement establishes the human-machine share control that is legally required on the use of a given weapon system (Amoroso & Tamburrini, 2021).

Based on the taxonomy of Sharkey (2016) the authors (Amoroso & Tamburrini, 2021, p. 261) propose five basic levels (L) of human-machine interactions to use as 'then' part of the bridge rules:

L1: A human engages with and selects targets, and initiates any attack.
L2: A program suggests alternative targets and a human chooses which to attack.
L3: A program selects targets and a human must approve them before the attack.
L4: A program selects and engages targets, but is supervised by a human who retains the power to override its choices and abort the attack.
L5: A program selects targets and initiates an attack on the basis of the mission goals as defined at the planning/activation stage, without further human involvement.

As an example, the following if-then rule is given: '"*IF the weapons system is programmed to perform an exclusively antimateriel defensive function (what property) AND is deployed in a sufficiently structured scenario (where property), THEN (L4) human operators must be put in charge of supervising the weapon's selection of targets and be given the power to override its choices*." (Amoroso & Tamburrini, 2021, p. 261).

Another approach to operationalise Meaningful Human Control is presented by Umbrello (2021) in which he couples two different Levels of Abstraction (LoA) to achieve Meaningful Human Control over an Autonomous Weapon System. In this he combines systems thinking and systems engineering as conceptual tools to frame the commonalities between these two LoAs. The author views a broader decision-making mechanism before deployment in which different agents have different levels of control over a specific part of the process. The concept of Meaning Human Control should reflect this and be positioned within the larger distributed network of decision-making. The systems thinking LoA helps conceptualizing the procedural processes such as operational planning and target identification while the systems engineering LoA aids with understanding both the tracing design as well as tracking the responsiveness of autonomous systems to the relevant moral reasons of the relevant agents. By this, it is possible to design for complex emergent behaviours and boundaries of systems. To achieve Meaningful Human Control, both LoA's are required and need to be coupled (Umbrello, 2021).

A third approach of operationalising Meaningful Human Control is that of Cavalcante Siebert et al. (2023) who are proposing four actional properties for AI-based systems under Meaningful Human Control to bridge the gap between philosophical theory and engineering practice. Building on the two necessary conditions for meaningful human control - tracking and tracing - distinct by Santoni de Sio & Van den Hoven (2018), the properties Cavalcante Siebert et al. (2023, p. 251) propose are:

Property 1: The human-AI system has an explicit moral operational design domain (moral ODD) and the AI agent adheres to the boundaries of this domain.

Property 2: Human and AI agents have appropriate and mutually compatible representations of the human-AI system and its context.

Property 3: The relevant agents have ability and authority to control the system so that humans can act upon their responsibility.

Property 4: Actions of the AI agents are explicitly linked to actions of humans who are aware of their moral responsibility.

In their reflection on their work the authors highlight that '*Meaningful human control is necessary but not sufficient for ethical AI.*' (Cavalcante Siebert et al., 2022, p. 252). The authors amplify this by stating that for a human-AI system to align with societal values and norms, Meaningful Human Control must entail a larger set design objectives which can be achieved by transdisciplinary practices.

All three approaches have in common that they focus on the human-machine interaction in order to operationalise the concept of Meaningful Human Control. Either by bridging the gap between weapon usage and ethical principles based on 'if-then' rules (Amoroso & Tamburrini, 2021), creating actional properties for the design of AI systems in which each of the properties human and artificial agents interact (Cavalcante Siebert et al., 2022), or proposing two LoA's in which different agents have different levels of control over the decision-making process to deploy an Autonomous Weapon System (Umbrello, 2021). Mirroring these three approaches to the Comprehensive Human Oversight Framework we can conclude that they can be positioned in the socio-technical layer of the framework which describes the human-machine interaction. This entails that all three approaches disregard the governance layer of the Comprehensive Human Oversight Framework which describes the supervision processes for Human Oversight. Therefore, we can conclude that, although valuable for operationalising the concept for Meaningful Human Control, these approaches do not provide new insights for Human Oversight of Autonomous Weapon Systems in order to ensure accountability and responsibility.

7.2 LIMITATIONS

The Comphrensive Human Oversight Framework introduced in this dissertation and the simulation of the implementation concept to operationalise the criteria for Human Oversight are new contributions of this research to the current academic work. Several limitations of these novelties can be identified. First of all, both the Comphrensive Human Oversight Framework as the simulation of the implementation concept are in this research only applied to one case: that of the deployment of an Aerial Autonomous Weapon Systems. The scenario and implementation concept are based on a tactical military context and therefore highlight specific tactical oversight issues. A scenario based on an operational or strategic military context might identify other oversight issues. The instruments described in the blocks of the Comphrensive Human Oversight Framework are generic and retrieved from a literature review from diverse academic disciplines. The simulation of the implementation concept is derived from the Glass Box framework. Both have not been applied to other case studies therefore it is not clear if they can be generalized to other application areas.

Secondly, due to time constraints we did not validate the Comphrensive Human Oversight Framework, the simulation of the implementation concept and Five-Point Systems Framework as we did with the results from the Value Deliberation Process. We did show the Comphrensive Human Oversight Framework to several experts, presented it to our peers, for example in several Doctoral Consortiums of conferences, and published both in peer-reviewed journal papers, but a structured rigorous scientific review is missing in this research.

The final limitation is that of the Value Deliberation process itself. The aim of the method is to get a mutual understanding of the different participant perspectives. However, the perspectives and opinions of the experts in the field of Autonomous Weapon Systems deployment are very distinct both pros and cons Autonomous Weapon Systems. During the value discussion, the participants gained more insight into each other's perspectives, but a value deliberation discussion will not lead to mutual understanding among the participants, as is originally aimed with this method, on the arguments pro or con the deployment of Autonomous Weapon Systems. As the debate on Autonomous Weapon Systems is polarized, the discussion on Autonomous Weapon Systems- mostly conducted in academic papers and in online blogs- is often one sided, e.g., either pros or cons, and a value deliberation can provide a balanced discussion on the topic and increase understanding.

7.3 FUTURE WORK

A few open questions remain for extending or adapting our work which can be divided in methodological and substantive improvements of our research. Methodologically, future work should address the limitations described in the previous section. The Comphrensive Human Oversight Framework and the implementation concept need to be applied to different case studies to see if they also applicable to other fields where autonomous systems are used. For example, in the case of Autonomous Vehicles and firefighting or humanitarian disaster relief with autonomous drones. It would be interesting to study which control instruments are used in these domains and to see if there are any control gaps that need to be filled for humans to remain in control and ensure accountability over these systems. Another methodological direction for future work is validating the Comphrensive Human Oversight Framework, the implementation concept and Five-Point Systems Framework with a rigorous scientific method to review and improve them if necessary.

Substantively, a research area for future work are the instruments in the blocks of the Comphrensive Human Oversight Framework. Especially the adding a governance instrument during the deployment phase is in block 2 is crucial. In the military domain the targeting process is an instrument that is used during the deployment of a weapon to govern its usage, but an oversight mechanism in block 2 seems to be missing which indicates a gap in the governance layer. As an oversight process is lacking, there is no sufficient mechanism for an institution to govern or supervise the ongoing control (block 5) of a (weapon) system in the socio-technical layer. The lack of an oversight mechanism in block 2 may lead to deficiencies in the ongoing control mechanism in block 5. A second substantive area for further research is the implementation concept as operationalisation of the Glass Box framework, which aims at monitoring the behaviour of autonomous systems during operations, for example decisions of a drone during its flight. In our implementation concept we do not monitor in-flight communication as we assume this is not possible in our scenario. In a future iteration of the implementation concept a scenario with in-flight communication should be studied to see how the safety of the system and its decisions should be monitored and documented in order to account for its behaviour if norm violations occur. Another direction for future work is extending the implementation concept to other values such as privacy (for example during information gathering nearby a village) and other human rights including more fuzzier norms. This will increase the usability of the implementation concept and enhance the accountability of the autonomous systems it monitors.

7.4 CONTRIBUTIONS

The scientific contributions of this research are twofold in that (1) our research contributes to a delineation of accountability, responsibility and Human Oversight that adds to the current body of literature, (2) insight into people's perception on accountability and responsibility during the deployment of an Autonomous Weapon Systems based on an empirical value elicitation process, and (3) the framework and implementation concept for Human Oversight for Autonomous Weapon Systems might also be applied to other fields to enhance transparency of decision-making by algorithms for Autonomous Systems, such as those for Autonomous Vehicles or in the medical domain.

The societal contribution of our research is a framework and implementation concept for Human Oversight that would lead to a proper allocation of accountability in the decision-making of the deployment of an Autonomous Weapon System and it might be possible to attribute responsibility for the actions taken by the weapon system by identifying the supervisor of these actions. This thereby contributes to decreasing the likelihood of unintended consequences in the deployment of Autonomous Weapon Systems.

7.5 POLICY RECOMMENDATIONS

In this section we collect the policy recommendations that we have made throughout this dissertation. By summarizing them, we provide an overview for policy makers and other stakeholders. Based on our research we would like to make several recommendations:

- The first relates to the introduction of autonomy in Autonomous Weapon Systems which has implications on the military control mechanisms, mainly in the socio-technical layer during deployment of an Autonomous Weapon System. This may require reformation of the military control instruments. These implications might lead to new training methods for military personnel for them to have the capacity (knowledge and skills) to responsibly deploy these weapons, but might also lead to new institutions and design methods, for example Value-Sensitive Design in military engineering, as control mechanisms in the governance layer.

- The second is that value deliberation using the Value Deliberation process could be beneficial in other areas than Autonomous Weapon Systems for policy making and citizen participation in local and national public administration. For example, to get citizen views on a municipal plan for the redevelopment of a local park or on a national level get input for nitrogen reduction policy. The application of the online Value Deliberation process method is not limited to Autonomous Weapon Systems and can be used in other areas as well.

- The third recommendation is regarding the feedback process of the Comphrensive Human Oversight Framework. The *monitoring*, *responding* and *acting* on the outcome of the review process can be done by an article 36 weapon review committee. Several countries have an article 36 reviewing procedure in place to conduct weapon reviews when new weapons and methods or means of warfare are studied, developed, acquired or adopted. The aim of the article 36 weapon review is to monitor the development of weapons by reference to its obligations under international humanitarian law by a State, but very few countries have a formal review mechanism in place and the format and responsibilities of the reviewing authority, how states interpret the terms of reference and legal obligations of Article 36 are conducted differently by each country. A standard process is lacking and developing an international standard article 36 review process could ensure the incorporation of the obligations and recommendations in the next iteration of the deployment of an Autonomous Weapon System. Sharing best practices of weapon reviews could be beneficial to improve the feedback process. Therefore, we recommend drafting an international standard article 36 review process and share the best practices among States. This would improve the allocation of accountability and responsibility in the deployment of Autonomous Weapon Systems.

Bibliography

List of publications

Summary

Samenvatting

Appendix A. Questionnaire Value Deliberation Process

Appendix B. Questionnaire validation Value Deliberation Process results

Appendix C. Overview key insights from literature review

Biography

BIBLIOGRAPHY

AIV, & CAVV. (2015). *Autonomous weapon systems: the need for meaningful human control*. (No. 97, No. 26). Retrieved from
https://www.advisorycouncilinternationalaffairs.nl/binaries/advisorycouncilinternationalaffairs/documents/publications/2015/10/02/autonomous-weapon-systems/Autonomous_Weapon_Systems_AIV-Advice-97_CAVV-Advisory-report-26_ENG_201510.pdf

Alfano, M., & Loeb, D. (2014). Experimental Moral Philosophy. Retrieved from https://plato.stanford.edu/entries/experimental-moral/

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

Aldewereld, H., Dignum, V., & Tan, Y.-h. (2015). Design for Values Information and communication technologies in Software Development Software development. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 831-845.

Alberts, D. S., & Hayes, R. E. (2006). *Understanding command and control* (pp. 31-33). Washington, DC: CCRP Publications.

Aler Tubella, A., & Dignum, V. (2019). *The Glass Box Approach: Verifying Contextual Adherence to Values.* Paper presented at the AISafety 2019, Macao, August 11-12, 2019.

Aler Tubella, A., Theodorou, A., Dignum, V., & Dignum, F. (2019). Governance by glass-box: implementing transparent moral bounds for AI behaviour. *arXiv preprint arXiv:1905.04994.*

Altmann, J., Asaro, P., Sharkey, N., & Sparrow, R. (2013). Armed military robots: editorial. *Ethics and Information Technology, 15*(2), 73.

Alston, P. (2010). *Study on Targeted Killings', Report of the UN Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, UN Doc A. United Nations. URL: https://www2.ohchr.org/english/bodies/hrcouncil/docs/14session/A. HRC*, *14*, 24.

Amoroso, D., & Tamburrini, G. (2021). Toward a Normative Model of Meaningful Human Control over Weapons Systems. *Ethics & International Affairs, 35*(2), 245-272.

Anderson, M., Anderson, S. L., & Berenz, V. (2016). *Ensuring Ethical Behavior from Autonomous Systems.* Paper presented at the AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments.

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE, 100*(3), 571-589.

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.

Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *Ai & Society, 35*, 611-623.

Åström, K. J., & Kumar, P. R. (2014). Control: A perspective. *Automatica, 50*(1), 3-43.

Beauchamp, T. L., & Walters, L. R. (1999). *Contemporary Issues in Bioethics*: Wadsworth Pub.

Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology, 20*(1), 41-58.

Borning, A., & Muller, M. (2012). *Next steps for value sensitive design.* Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.

Bovens, M., Schillemans, T., & Goodin, R. E. (2014). Public accountability. *The Oxford handbook of public accountability, 1*, 1-20.

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. *European law journal, 13*(4), 447-468.

Bovens, M. (2014). Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism. In *Accountability and European governance* (pp. 28-49): Routledge.

Bradley, B. (2006). Two concepts of intrinsic value. *Ethical Theory and Moral Practice, 9*(2), 111-130.

Bradshaw, J. M., Hoffman, R. R., Woods, D. D., & Johnson, M. (2013). The seven deadly myths of" autonomous systems". *IEEE intelligent systems, 28*(3), 54-61.

Broeks, G., van den Herik, L., Aerdts, W., Casteleijn, L., Ginkel, B. v., Groot, J. d., . . . Ryngaert, C. (2021). Autonome wapensystemen: het belang van reguleren en investeren. *Autonome wapensystemen: het belang van reguleren en investeren*.

Busuioc, M. (2007). *Autonomy, Accountability and Control. The Case of European Agencies.* Paper presented at the 4TH ECPR General Conference, Pisa, Italy.

Campaign Stop Killer Robots. (2023). Problems with Autonomous Weapons. Stop Killer Robots. Retrieved from https://www.stopkillerrobots.org/stop-killer-robots/facts-about-autonomous-weapons/.

Caparini, M. (2004). Media and the security sector: Oversight and accountability. *Geneva Centre for the Democratic Control of Armed Forces (DCAF) Publication*, 1-49.

Castelfranchi, C. (1995). Intelligence Agents: Thories, Architectures, and Languages. *Guarantees for autonomy in cognitive agent architecture, (edited by M. Wooldridge and NR Jennings), Springer*.

Castelfranchi, C., & Falcone, R. (2003). From automaticity to autonomy: the frontier of artificial agents. In *Agent Autonomy* (pp. 103-136): Springer.

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., . . . Jonker, C. M. (2023). Meaningful human control: actionable properties for AI system development. *AI and Ethics, 3*(1), 241-255.

Cheng, A. S., & Fleischmann, K. R. (2010). Developing a meta-inventory of human values. *Proceedings of the American Society for Information Science and Technology, 47*(1), 1-10.

Clarke, R., & Moses, L. B. (2014). The regulation of civilian drones' impacts on public safety. *Computer law & security review, 30*(3), 263-285.

Collingridge, D. (1980). The social control of technology. *Frances Pinter*, London

Cordeschi, R. (2013). Automatic decision-making and reliability in robotic systems: some implications in the case of robot weapons. *Ai & Society, 28*(4), 431-441.

Cornwall, A. (2008). "Unpacking Participation: Models, Meanings, and Practices." Community

Development Journal 43 (3): 269–83.

Côté, N., Bouzid, M., & Mouaddib, A.-I. (2011). *Integrating the human recommendations in the decision process of autonomous agents: A goal biased markov decision process.* Paper presented at the Proceedings of the AAAI 2011 Fall Symposium Robot-Human Teamwork in Dynamic Adverse Environmen.

Crootof, R. (2015). War torts: Accountability for autonomous weapons. *U. Pa. L. Rev., 164*, 1347.

Cummings, M. L. (2006). Integrating ethics in design through the value-sensitive design approach. *Science and engineering ethics, 12*(4), 701-715.

Cummings, M. L. (2006a). Automation and accountability in decision support system interface design. Retrieved from https://dspace.mit.edu/handle/1721.1/90321

Cummings, M. L. (2006b). Integrating ethics in design through the value-sensitive design approach. *Science and engineering ethics, 12*(4), 701-715.

Davis, J., & Nathan, L. P. (2015). Value Sensitive Design: Applications, Adaptations, and Critiques. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 11-40.

Dawes, J. (2023). Killer robots are the future of warfare and the 'inevitable next step' in Russia's long bloody invasion of Ukraine. Retrieved from https://fortune.com/2023/02/21/killer-robots-a-i-future-warfare-russia-ukraine-invasion/

de Ágreda, Á. G. (2020). Ethics of autonomous weapons systems and its applicability to any AI systems. *Telecommunications Policy, 44*(6), 101953.

Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems, 77*, 1-14.

De Wet, E. (2008). Holding international institutions accountable: the complementary role of non-judicial oversight mechanisms and judicial review. *German Law Journal, 9*(11), 1987-2012.

Dickinson, L. (2018). Lethal Autonomous Weapons Systems: The Overlooked Importance of Administrative Accountability. *Lethal Autonomous Weapons Systems: The overlooked importance of administrative accountability, in The Impact of Emerging Technologies on the Law of Armed Conflict (Eric Talbot Jensen & Ronald Alcala eds., Oxford University Press 2018 Forthcoming).*

Docherty, B. (2012). *Losing humanity: The case against killer robots*. *Human Right Watch, https://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf*

Docherty, B. (2015). *Mind the gap: The lack of accountability for killer robots*: Human Rights Watch.

Dryzek, J. S., & Niemeyer, S. (2006). Reconciling pluralism and consensus as political ideals. *American Journal of Political Science, 50*(3), 634-649.

Ekelhof, M. (2015). Autonome wapens: een verkenning van het concept Meaningful Human Control. *Militaire Spectator, 184*.

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

European Commission. (2019). *Ethics Guidelines for Trustworthy Artificial Intelligence*. Retrieved from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Fischer, J. M., & Ravizza, M. (1998). Responsibility and Control: A Theory of Moral Responsibility.

Fishkin, J. (2009). *When the people speak: Deliberative democracy and public consultation*: Oup Oxford.

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

Friedman, B., & Kahn Jr, P. H. (2003). Human values, ethics, and design. *The human-computer interaction handbook*, 1177-1201.

Friedman, B., Kahn Jr, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory* (pp. 55-95): Springer.

Galliott, J. (2015). *Military robots: Mapping the moral landscape*: Ashgate Publishing, Ltd.

Gardner, J. (2007). The Mark of Responsibility. In *Offences and Defences* (pp. 177–200): Oxford University Press.

General Assembly United Nations. (2016). *Joint report of the Special Rapporteur on the rights to freedom of peaceful* *assembly and of association and the Special Rapporteur on extrajudicial, summary or arbitrary executions on the proper management of assemblies*. (A/HRC/31/66).

Genesereth, M., & Ketchpel, S. (1994). Association for computing machinery. *Software Agents, 37*(7), 48-59.

Gigler, B.-S., Custer, S., Bailur, S., Dodds, E., Asad, S., & Gagieva-Petrova, E. (2014). Closing the feedback loop: Can technology amplify citizen voices. *Closing the Feedback Loop*, 211.

Goodin, R. E. (1995). *Utilitarianism as a public philosophy*: Cambridge University Press.

Gouveia, V. V., Milfont, T. L., & Guerra, V. M. (2014). Functional theory of human values: Testing its content and structure hypotheses. *Personality and Individual Differences, 60*, 41-47.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2012). Moral foundations theory: The pragmatic validity of moral pluralism.

Greer, S. L., Wismar, M., Figueras, J., & McKee, C. (2016). Governance: a framework. *Strengthening Health System Governance*, 27-56.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus, 133*(4), 55-66.

Hartmann, S. (1996). The world as a process: Simulations in the natural and social sciences. In *Modelling and simulation in the social sciences from the philosophy of science point of view* (pp. 77-100): Springer.

Horowitz, M., & Scharre, P. (2015). *Meaningful human control in weapon systems: a primer*: Center for a New American Security.

Horowitz, M. C., & Scharre, P. (2015). MEANINGFUL HUMAN CONTROL in WEAPON SYSTEMS. Retrieved from https://www.jstor.org/stable/pdf/resrep06179.pdf

Horowitz, M. C. (2016). The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons. *Daedalus, 145*(4), 25-36.

Hulstijn, J., & Burgemeestre, B. (2014). Design for the Values of Accountability and Transparency. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 1-25.

Human Rights Watch. (2023). Statement to Convention on Conventional Weapons GGE Meeting on Lethal Autonomous Weapons System [Press release]. Retrieved from https://www.hrw.org/news/2023/03/06/statement-convention-conventional-weapons-gge-meeting-lethal-autonomous-weapons

HRW, & IHRC. (2012). *Losing Humanity: The Case against Killer Robots*. Retrieved from https://www.hrw.org/sites/default/files/reports/arms1112ForUpload_0_0.pdf

Hurka, T. (2005). Proportionality in the Morality of War. *Philosophy & Public Affairs, 33*(1), 34-66.

ICRC, I. C. o. t. R. C. (2010, 29-10-2010). War and international humanitarian law. Retrieved from https://www.icrc.org/eng/war-and-law/overview-war-and-law.htm

ICRC. (2023). ICRC: "Autonomous weapons represent an urgent humanitarian priority today" [Press release]. Retrieved from https://www.icrc.org/en/document/icrc-autonomous-weapons-represent-urgent-humanitarian-priority-today-0

Jacobs, A. (2010). Creating the missing feedback loop. *IDS Bulletin, 41*(6), 56-64.

Jensen, K. (1994). *An introduction to the theoretical aspects of coloured petri nets.* Paper presented at the A Decade of Concurrency Reflections and Perspectives: REX School/Symposium Noordwijkerhout, The Netherlands June 1–4, 1993 Proceedings.

Jensen, K., Kristensen, L. M., & Wells, L. (2007). Coloured Petri Nets and CPN Tools for modelling and validation of concurrent systems. *International Journal on Software Tools for Technology Transfer, 9*(3), 213-254.

Johnson, A. M., & Axinn, S. (2013). The morality of autonomous robots. *Journal of Military Ethics, 12*(2), 129-141.

JUAS-COE. (2010). *JFCOM-UAS-PocketGuide*. Arlington County VA USA: JUAS-COE

Kaag, J., & Kaufman, W. (2009). Military frameworks: Technological know-how and the legitimization of warfare. *Cambridge Review of International Affairs, 22*(4), 585-606.

Kheir, N., Åström, K. J., Auslander, D., Cheok, K. C., Franklin, G. F., Masten, M., & Rabins, M. (1996). Control systems engineering education. *Automatica, 32*(2), 147-166.

Keohane, R. O. (2003). *Global governance and democratic accountability*: Citeseer.

Koppell, J. G. (2005). Pathologies of accountability: ICANN and the challenge of "multiple accountabilities disorder". *Public administration review, 65*(1), 94-108.

Kramer, M. F., Borg, J. S., Conitzer, V., & Sinnott-Armstrong, W. (2017). When Do People Want AI to Make Decisions?

Kroes, P., & van de Poel, I. (2015). Design for Values and the Definition Value measurement, Specification Value specification, and Operationalization Value operationalization of Values. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 151-178.

Kuptel, A., & Williams, A. (2014). Policy Guidance: Autonomy in Defence Systems. *Available at SSRN 2524515*.

Li, S.-M., Boskovic, J. D., Seereeram, S., Prasanth, R., Amin, J., Mehra, R. K., . . . McLain, T. W. (2002). *Autonomous hierachical control of multiple unmanned combat air vehicles (UCAVs).* Paper presented at the Proceedings of the American control conference.

Liao, S.-H. (2008). Problem structuring methods in military command and control. *Expert Systems with applications, 35*(3), 645-653.

Lindner, F., Bentzen, M. M., & Nebel, B. (2017, September). The HERA approach to morally competent robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 6991-6997). IEEE.

Luppicini, R., & So, A. (2016). A technoethical review of commercial drone use in the context of governance, ethics, and privacy. *Technology in society, 46*, 109-119.

NATO. (2016). *ALLIED JOINT DOCTRINE FOR JOINT TARGETING Edition A Version 1* Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/628215/20160505-nato_targeting_ajp_3_9.pdf

NATO. (2017). *AJP-01 ALLIED JOINT DOCTRINE*. Retrieved from

https://www.gov.uk/government/publications/ajp-01-d-allied-joint-doctrine

Mohan, G. (2001). "Participatory Development." In The Arnold Companion to Development

Studies, edited by V. Desai and R. Potter, 49–54. London: Hodder Arnold.

Maslow, A. H. (1943). A theory of human motivation. *Psychological review, 50*(4), 370.

Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: guided by information power. *Qualitative health research, 26*(13), 1753-1760.

Marchant, G. E., Allenby, B., Arkin, R. C., Barrett, E. T., Borenstein, J., Gaudet, L. M., . . . O'Meara, R. (2011). International governance of autonomous military robots.

Marshall, B., Cardon, P., Poddar, A., & Fontenot, R. (2013). Does sample size matter in qualitative research?: A review of qualitative interviews in IS research. *Journal of computer information systems, 54*(1), 11-22.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175-183.

McClelland, J. (2003). The review of weapons in accordance with Article 36 of Additional Protocol I. *International Review of the Red Cross, 85*(850), 397-420.

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

Melancon, A.-A. (2020). What's wrong with drones? Automatization and target selection. *Small Wars & Insurgencies, 31*(4), 801-821.

Meloni, C. (2016). State and Individual Responsibility for Targeted Killings by Drones. *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on Remotely Controlled Weapons*, 47.

Miller, K. W., Wolf, M. J., & Grodzinsky, F. (2017). This "ethical trap" is for roboticists, not robots: on the issue of artificial agent ethical decision-making. *Science and engineering ethics, 23*(2), 389-401.

Mulgan, R. (2000). 'Accountability': An ever-expanding concept? *Public administration, 78*(3), 555-573.

Open Roboethics initiative. (2015, 5-11-2015). The Ethics and Governance of Lethal Autonomous Weapons Systems: An International Public Opinion Poll. Retrieved from http://www.openroboethics.org/laws_survey_released/

Pelizzo, R., Stapenhurst, R., & Olson, D. (2006). Parliamentary oversight for government accountability. Retrieved from

https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=1136&context=soss_research

Pesch, U. (2015). Engineers and active responsibility. *Science and engineering ethics, 21*(4), 925-939.

Pigeau, R., & McCann, C. (2002). *Re-conceptualizing command and control*. Defence R & D Canada-Toronto.

Pigmans, K. (2020). Value Deliberation: Towards mutual understanding of stakeholder perspectives in policymaking. Retrieved from https://research.tudelft.nl/en/publications/value-deliberation-towards-mutual-understanding-of-stakeholder-pe

Pigmans, K., Dignum, V., & Doorn, N. (2021). Group proximity and mutual understanding: measuring onsite impact of a citizens' summit. *Journal of Public Policy, 41*(2), 228-250.

Ricard, M., & Kolitz, S. (2003). *The ADEPT framework for intelligent autonomy*. Charles Stark Draper Laboratory, Incorporated.

Roff, H. M. (2013). Responsibility, liability, and lethal autonomous robots. *Routledge Handbook of Ethics and War: Just War Theory in the 21st Century. Routledge*, 352-364.

Romzek, B. S., & Dubnick, M. J. (1987). Accountability in the public sector: Lessons from the Challenger tragedy. *Public administration review*, 227-238.

Radin, B. A., & Romzek, B. S. (1996). Accountability expectations in an intergovernmental arena: The national rural development partnership. *Publius: The Journal of Federalism, 26*(2), 59-81.

Rao, B., Gopi, A. G., & Maione, R. (2016). The societal impact of commercial drones. *Technology in society, 45*, 83-90.

Roff, H. M. (2013). *Responsibility, liability, and lethal autonomous robots* (pp. 352–364)., Routledge handbook of ethics and war: just war theory in the 21st century Abingdon: Routledge.

Roff, H. M. (2016). Weapons autonomy is rocketing. Retrieved from

http://foreignpolicy.com/2016/09/28/weapons-autonomy-is-rocketing/

Roff, H. M., & Moyes, R. (2016). *Meaningful human control, artificial intelligence and autonomous weapons.* Paper presented at the Briefing Paper Prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons.

Rosén, F. (2014). Extremely stealthy and incredibly close: drones, control and legal responsibility. *Journal of Conflict and Security Law, 19*(1), 113-131.

Rosenberg, M., & Markoff, J. (2016). The Pentagon's 'Terminator Conundrum': Robots That Could Kill on Their Own. *The New York Times*. Retrieved from

http://www.nytimes.com/2016/10/26/us/pentagon-artificial-intelligence-terminator.html?_r=0

Royakkers, L., & Orbons, S. (2015). Design for Values in the Armed Forces: Nonlethal Weapons Weapons and Military

Military Robots Robot. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 613-638.

Rønnow-Rasmussen, T. (2002). Instrumental values–Strong and weak. *Ethical Theory and Moral Practice, 5*(1), 23-43.

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI, 5*, 15.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology, 34*, 1057-1084.

Sayler, K. M. (2021). *International discussions concerning lethal autonomous weapon systems*.

Schedler, A. (1999). Conceptualizing accountability*. The self-restraining state: Power and accountability in new democracies, 14*.

Schroeder, M. (2016). Value Theory. Retrieved from https://plato.stanford.edu/entries/value-theory/

Scott, C. (2000). Accountability in the regulatory state. *Journal of law and society, 27*(1), 38-60.

Searle, J. R. (1995). *The construction of social reality*. Simon and Schuster.

Shapiro, R. M., Pinci, V. O., & Mameli, R. (1993). Modeling a NORAD command post using SADT and colored Petri nets. *Functional Programming, Concurrency, Simulation and Automated Reasoning: International Lecture Series 1991–1992 McMaster University, Hamilton, Ontario, Canada*, 84-107.

Sharkey, N., & Suchman, L. (2013). *Wishful mnemonics and autonomous killing machines*. Paper presented at the Proceedings of the AISB.

Soldak, K. (2023). Friday, January 27. Russia's War On Ukraine: Daily News And Information From Ukraine. Retrieved from https://www.forbes.com/sites/katyasoldak/2023/01/27/friday-january-27-russias-war-on-ukraine-daily-news-and-information-from-ukraine/?sh=137c4bf422b4

Sparrow, R. (2016). Robots and respect: Assessing the case against autonomous weapon systems. *Ethics & International Affairs, 30*(1), 93-116.

Schwarz, E. (2018). The (im)possibility of meaningful human control for lethal autonomous weapon systems. Retrieved from http://blogs.icrc.org/law-and-policy/2018/08/29/im-possibility-meaningful-human-control-lethal-autonomous-weapon-systems/

Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of social issues, 50*(4), 19-45.

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture, 2*(1), 11.

Sharkey, N. (2016). Staying in the loop: human supervisory control of weapons. *Autonomous weapons systems: Law, ethics, policy*, 23-38.

Tenny, S., Brannan, G. D., Brannan, J. M., & Sharts-Hopko, N. C. (2022). Qualitative Study. [Internet]. *StatPearls [Internet]. StatPearls Publishing.*

Taddeo, M., & Blanchard, A. (2022). A comparative analysis of the definitions of autonomous weapons systems. *Science and engineering ethics, 28*(5), 37.

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

Tucker, P. (2023). Russian Robot Maker Working On Bot to Target Abrams, Leopard Tanks Retrieved from https://www.defenseone.com/technology/2023/01/russian-robot-maker-working-bot-target-abrams-leopard-tanks/382288/

UN GGE LAWS. (2018). Emerging Commonalities, Conclusions and Recommendations. https ://www.unog.ch/80256EDD00 6B895 4/(httpA ssets)/EB4EC 9367D 3B63B 1C125 82FD0 057A9 A4/$file/GGE+LAWS+Augus t_EC,+C+and+Rs_final .pdf

UNDIR. (2014). *Framing Discussions on the Weaponization of Increasingly Autonomous Technologies*. Retrieved from http://www.unidir.org/files/publications/pdfs/framing-discussions-on-the-weaponization-of-increasingly-autonomous-technologies-en-606.pdf

UNDIR. (2015). *The Weaponization of Increasingly Autonomous Technologies: Considering Ethics and Social Values*. Retrieved from http://www.unidir.org/files/publications/pdfs/considering-ethics-and-social-values-en-624.pdf.

Umbrello, S. (2020). Meaningful human control over smart home systems: a value sensitive design approach. *HUMANA. MENTE, 13*(37), 40-65.

Umbrello, S., & Van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics, 1*(3), 283-296.

Umbrello, S. (2021). Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: A two-tiered approach. *Ethics and Information Technology, 23*(3), 455-464.

US Department of Defense. (2023). DoD Directive 3000.09: Autonomy in Weapon Systems. Retrieved from: https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

Vas, E., Lescroël, A., Duriez, O., Boguszewski, G., & Grémillet, D. (2015). Approaching birds with drones: first experiments and ethical guidelines. *Biology letters, 11*(2), 20140754.

Van den Berg, J. (2015). Wat maakt cyber security anders dan informatiebeveiliging? *Magazine Nationale Veiligheid en Crisisbeheersing,(2) 2015*.

van den Hoven, J., Vermaas, P. E., & van de Poel, I. (2015). Design for values: An introduction. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 1-7.

Van de Poel, I. (2011). The relation between forward-looking and backward-looking responsibility. In *Moral responsibility* (pp. 37-52): Springer.

Van de Poel, I. (2013). Translating values into design requirements. In *Philosophy and engineering: Reflections on practice, principles and process* (pp. 253-266): Springer.

van der Vecht, B. (2009). *Adjustable autonomy: Controling influences on decision making* (Doctoral dissertation, University Utrecht).

van Wynsberghe, A., & Robbins, S. (2014). Ethicist as Designer: a pragmatic approach to ethics in the lab. *Science and engineering ethics, 20*(4), 947-961.

van Wynsberghe, A., & Comes, T. (2020). Drones in humanitarian contexts, robot ethics, and the human–robot interaction. *Ethics and Information Technology, 22*(1), 43-53.

Vanderelst, D., & Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research, 48*, 56-66.

Verbruggen, M., & Boulanin, V. (2017). SIPRI Compendium on Article 36 Reviews. Retrieved from https://researchportal.vub.be/en/publications/sipri-compendium-on-article-36-reviews

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

Verdiesen, I. (2017). *Agency perception and moral values related to Autonomous Weapons: An empirical study using the Value-Sensitive Design approach.*

Verdiesen, I., Aler Tubella, A., & Dignum, V. (2021). Integrating Comprehensive Human Oversight in Drone Deployment: A Conceptual Framework Applied to the Case of Military Surveillance Drones. *Information, 12*(9), 385.

Verdiesen, I., & Dignum, V. (2022). Value elicitation on a scenario of autonomous weapon system deployment: a qualitative study based on the value deliberation process. *AI and Ethics*, 1-14.

Verdiesen, I., Dignum, V., & Hoven, J. V. D. (2018). Measuring moral acceptability in E-deliberation: A practical application of ethics by participation. *ACM Transactions on Internet Technology (TOIT), 18*(4), 1-20.

Verdiesen, I., De Sio, F. S., & Dignum, V. (2019). Moral Values Related to Autonomous Weapon Systems: An Empirical Survey that Reveals Common Ground for the Ethical Debate. *IEEE Technology and Society Magazine, 38*(4), 34-44.

Vignard, K. (2014). The weaponization of increasingly autonomous technologies: considering how meaningful human control might move discussion forward. *UNIDIR Resources, 2*.

Wagner, M. (2014). The dehumanization of international humanitarian law: Legal, ethical, and political
implications of autonomous weapon systems. *Vanderbilt Journal of Transnational Law, 47, 1371.*

Walsh, J. I., & Schulzke, M. (2015). *The Ethics of Drone Strikes: Does Reducing the Cost of Conflict Encourage War?* Retrieved from

Westergaard, M., Evangelista, S., & Kristensen, L. M. (2009). *ASAP: an extensible platform for state space analysis.* Paper presented at the International Conference on Applications and Theory of Petri Nets.

Williams, A. P., Scharre, P. D., & Mayer, C. (2015). Developing Autonomous Systems in an Ethical Manner. In *Autonomous Systems: Issues for Defence Policymakers*: NATO Allied Command Transformation (Capability Engineering and Innovation).

Williams, G. (2008). Responsibility as a virtue. *Ethical Theory and Moral Practice, 11*(4), 455-470.

Williams, J. (2015). Democracy and regulating autonomous weapons: biting the bullet while missing the point? *Global Policy, 6*(3), 179-189.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The knowledge engineering review, 10*(2), 115-152.

Wright, G. H. v. (1981). On the logic of norms and actions. In *New studies in deontic logic* (pp. 3-35): Springer.

LIST OF PUBLICATIONS

Journal Papers

Verdiesen, I., & Dignum, V. (2023). Value elicitation on a scenario of autonomous weapon system deployment: a qualitative study based on the value deliberation process. *AI and Ethics*, *3*(3), 887-900.

Verdiesen, I., Aler Tubella, A., & Dignum, V. (2021). Integrating comprehensive human oversight in drone deployment: a conceptual framework applied to the case of military surveillance drones. *Information*, *12*(9), 385.

Verdiesen, I., Santoni de Sio, F., & Dignum, V. (2021). Accountability and control over autonomous weapon systems: a framework for comprehensive human oversight. *Minds and Machines*, *31*(1), 137-163.

Verdiesen, I., De Sio, F. S., & Dignum, V. (2019). Moral values related to autonomous weapon systems: an empirical survey that reveals common ground for the ethical debate. *IEEE Technology and Society Magazine*, *38*(4), 34-44.

Verdiesen, I., Dignum, V., & Hoven, J. V. D. (2018). Measuring moral acceptability in E-deliberation: A practical application of ethics by participation. *ACM Transactions on Internet Technology (TOIT)*, *18*(4), 1-20.

Workshop papers

Verdiesen, I. (2022, July). Comprehensive Human Oversight Framework to Ensure Accountability over Autonomous Weapon Systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 917-917).

Verdiesen, I. (2018, December). The design of human oversight in autonomous weapon systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 388-389).

Verdiesen, I., Dignum, V., & Rahwan, I. (2018). Design requirements for a moral machine for Autonomous Weapons. In *Computer Safety, Reliability, and Security: SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings 37* (pp. 494-506). Springer International Publishing.

Verdiesen, I., Cligge, M., Timmermans, J., Segers, L., Dignum, V., & van den Hoven, J. (2016). MOOD: Massive Open Online Deliberation Platform-A Practical Application. In *EDIA@ ECAI* (pp. 4-9).

Essay

Verdiesen, I. (2017). How do we ensure that we remain in control of our autonomous weapons?. *AI Matters*, *3*(3), 47-55.

Master thesis

Verdiesen, I. (2017). Agency perception and moral values related to Autonomous Weapons: An empirical study using the Value-Sensitive Design approach. https://repository.tudelft.nl/islandora/object/uuid:7cc28c2e-69d9-45f3-9c87-51e8281c32b0

Verdiesen, I. (2015). Innovation on demand: Can Innovative Work Behaviour be stimulated by High Performance Work Systems in a learning organisation? http://hdl.handle.net/1820/7911

SUMMARY

Autonomous Weapon Systems are weapons systems equipped with Artificial Intelligence (AI). They are increasingly deployed on the battlefield. Autonomous systems can have many benefits in the military domain, yet the nature of Autonomous Weapon Systems might also lead to security risks and unpredictable activities. Next to this, the lack of human dignity, which is linked to life-or-death decision-making, is mentioned as concern with the use of Autonomous Weapon Systems. At the same time, many scholars express concerns that Autonomous Weapon Systems will lead to an "accountability gap"; circumstances in which no human can be held responsible and accountable for the decisions, actions and effects of Autonomous Weapon Systems.

The aforementioned concerns display that responsibility, accountability and human control are values often mentioned in the societal and academic debate on Autonomous Weapon Systems. To the best of our knowledge, empirical studies on the extent how responsibility and accountability of the deployment of Autonomous Weapon Systems are perceived by common people and experts are missing. The notion of "Meaningful Human Control" is often mentioned as a condition in the debate on Autonomous Weapon Systems to ensure accountability and responsibility over these type of weapon systems. In our opinion, Meaningful Human Control alone will not suffice as requirement to minimize unintended consequences of Autonomous Weapon Systems due to several reasons. Firstly, the concept of Meaningful Human Control is potentially controversial and confusing as human control is defined and understood differently in various literature domains. Secondly, standard concepts of control in engineering and the military domain entail a capacity to directly cause or prevent an outcome that is not possible to achieve with an Autonomous Weapon System, because once an autonomous weapon is launched you cannot intervene by human action. And finally, specific literature on Meaningful Human Control over Autonomous Weapon Systems does not offer a consistent usable concept. We believe that a different approach is needed to minimize unintended consequences of Autonomous Weapons Systems. Therefore, we propose to rather focus on human oversight instead of Meaningful Human Control. This leads to the following research objective:

> *To improve the allocation of accountability and responsibility by designing a framework and implementation concept such that criteria for Human Oversight are identified, represented and validated in order to minimize unintended consequences in the deployment of Autonomous Weapon Systems.*

To achieve this research objective, we applied the Value-Sensitive Design (VSD) method

as research approach. The VSD is a three-partite approach that allows for considering human values throughout the design process of technology. It is an iterative process for the conceptual, empirical and technological investigation of human values implicated by the design of an artifact.

In the conceptual investigation phase of our research, we propose a Comprehensive Human Oversight Framework for Autonomous Weapon Systems. This framework consists of three layers that connect the technical, socio-technical and governance perspectives of control. These layers link the control perspectives to a time perspective which shows when a process is taking place with respect to autonomous action: (1) before deployment of a weapon, (2) during deployment of a weapon and (3) after deployment of a weapon. This results in a nine-block framework that contains several control mechanisms. Our main claim is that combining the control mechanisms in the technical, socio-technical and governance layer will lead to Comprehensive Human Oversight over Autonomous Weapon Systems which may ensure solid controllability and accountability for the behaviour of Autonomous Weapon Systems. When applied to the case of Autonomous Weapon Systems the Comprehensive Human Oversight Framework reveals two gaps in control, one gap in the governance layer and one in the socio-technical layer during deployment of an Autonomous Weapon System. The application of the Glass Box framework on the Comprehensive Human Oversight Framework could mitigate these gaps in control. The Glass Box framework is built around the black box (in the socio-technical layer during deployment of an Autonomous Weapon Systems) with an interpretation and the observation stage which allows for a transparent human oversight process. A feedback process can close the loop after deployment of a weapon from the accountability process back to the interpretation stage before a next deployment of a weapon.

As part of the empirical investigation phase of our research, we build on our conceptual work by conducting qualitative research through interviews, value deliberation in expert panels and a survey. We conducted value elicitation by means of the Value Deliberation Process as first step of the interpretation stage of the Glass Box framework. We designed an online Value Deliberation Process consisting of a bipartite survey and a virtual session for the expert panel discussion. The value deliberation was done with 14 participants divided over two groups. The participants were a mix of military personnel and civilians working at the Dutch Ministry of Defense, an NGO, researchers, policymakers and industry. The value elicitation conducted using the Value Deliberation Process not only shows that value discussion leads to changes in perception of the acceptability of alternatives in a scenario of Autonomous Weapon System deployment, it also gives insight into which values are deemed important and highlights that trust in the decision-making of an Autonomous Weapon System is crucial.

During the technical investigation phase of our research, we operationalised the Glass Box framework by creating an implementation concept as an example to prove that the framework is actionable. After value elicitation, deriving norms and requirements is the next step in the interpretation stage of the Glass Box framework. These requirements will feed into the observation stage as observable elements (criteria) to monitor and verify. We created an implementation of a pre- and post-flight procedure of an autonomous drone as an example using Coloured Petri Nets (CPNs) as modelling language. The implementation concept shows that it is possible to set criteria in the pre-flight process and to evaluate these criteria post-flight. During flight, the drone itself is treated as a black box of which the internal logic is not accessible. This way the users do not need technical skills to understand the internal workings of an autonomous drone, but still can monitor and oversee the use of the autonomous system based on observable norms. A review stage is required after deployment as an accountability process of which findings should feed back into the interpretation stage for a next deployment of an autonomous system and thereby close the loop between the stages.

Finally, we returned to the conceptual investigation phase of our research and describe a toy example in which we apply the Five-Point Systems feedback loop to the case of Autonomous Weapon Systems. The purpose of the feedback system is to ensure that the lessons and recommendations from the review stage will be incorporated in the interpretation stage before deployment of an Autonomous Weapon System in a next iteration.

Results of this research are the delineation of accountability, responsibility and human oversight concepts which adds to the current body of literature. Also, the Comprehensive Human Oversight Framework and implementation concept of the Glass Box framework show that criteria for Human Oversight can be identified, represented and validated. This leads to a proper allocation of accountability in the decision-making of the deployment of an Autonomous Weapon System and it might be possible to attribute responsibility for the actions taken by the weapon system by identifying the supervisor of these actions. This thereby contributes to decreasing the likelihood of unintended consequences in the deployment of Autonomous Weapon Systems.

SAMENVATTING

Autonome Wapen Systemen zijn wapensystemen die met Kunstmatige Intelligentie (KI) zijn uitgerust. Ze worden steeds vaker ingezet op het slagveld. Autonome systemen kunnen veel voordelen hebben in het militaire domein, maar de aard van Autonome Wapen Systemen kan ook leiden tot veiligheidsrisico's en onvoorspelbare activiteiten. Tevens wordt het gebrek aan menselijke waardigheid, dat verband houdt met besluitvorming over leven of dood, genoemd als een punt van zorg bij het gebruik van Autonome Wapen Systemen. Daarnaast zijn veel wetenschappers bezorgd dat Autonome Wapen Systemen zullen leiden tot een "verantwoordelijkheidshiaat"; omstandigheden waarin geen mens verantwoordelijk kan worden gehouden en aansprakelijk kan worden gesteld voor de beslissingen, acties en effecten van Autonome Wapen Systemen.

De bovengenoemde zorgen tonen aan dat verantwoordelijkheid, verantwoording en menselijke controle waarden zijn die vaak worden genoemd in het maatschappelijke en academische debat over Autonome Wapen Systemen. Voor zover wij weten, ontbreken empirische studies over de mate waarin verantwoordelijkheid en verantwoording voor de inzet van autonome wapensystemen door gewone mensen en experts worden ervaren. Het begrip "Betekenisvolle Menselijke Controle" wordt vaak genoemd als voorwaarde in het debat over Autonome Wapen Systemen om verantwoording en verantwoordelijkheid over dit soort wapensystemen te waarborgen. Naar onze mening is om verschillende redenen alleen Betekenisvolle Menselijke Controle niet voldoende als vereiste om onbedoelde gevolgen van Autonome Wapen Systemen te minimaliseren. Ten eerste is het concept van Betekenisvolle Menselijke Controle potentieel controversieel en verwarrend, aangezien menselijke controle in verschillende literatuurdomeinen verschillend wordt gedefinieerd en geintrepeteerd. Ten tweede houden standaardconcepten van controle in het technische en het militaire domein een vermogen in om direct een uitkomst te veroorzaken of te voorkomen die niet mogelijk is met een Autonome Wapen Systeem, omdat als een autonoom wapen eenmaal is gelanceerd, je niet kunt ingrijpen door menselijk handelen. En tot slot biedt de literatuur over Betekenisvolle Menselijke Controle over Autonome Wapen Systemen geen consistent bruikbaar concept. Wij zijn van mening dat er een andere aanpak nodig is om de onbedoelde gevolgen van Autonome Wapen Systemen tot een minimum te beperken. Daarom leggen we de focus op menselijk toezicht in plaats van op Betekenisvolle Menselijke Controle. Dit leidt tot het volgende onderzoeksdoel:

> *Het verbeteren van het afleggen van verantwoording en toewijzen van verantwoordelijkheid door een raamwerk en implementatieconcept te ontwerpen waarin criteria voor menselijk toezicht worden geïdentificeerd, weergegeven en gevalideerd om onbedoelde gevolgen bij de inzet van*

Autonome Wapen Systemen tot een minimum te beperken.

We hebben de Value-Sensitive Design (VSD) methode als onderzoeksaanpak toegepast om dit onderzoeksdoel te bereiken. Het VSD is een drieledige benadering die het mogelijk maakt om tijdens het hele ontwerpproces van technologie rekening te houden met menselijke waarden. Het is een iteratief proces voor het conceptuele, empirische en technologische onderzoek van menselijke waarden die betrekking hebben op het ontwerp van een artefact.

In de conceptuele onderzoeksfase stellen we een alomvattend menselijk toezichtskader - het Comprehensive Human Oversight Framework - voor Autonome Wapen Systemen voor. Dit raamwerk bestaat uit drie lagen die de technische, socio-technische en bestuurlijke controle perspectieven met elkaar verbinden. Deze lagen koppelen de controleperspectieven aan een tijdsperspectief dat laat zien wanneer een proces plaatsvindt: (1) voor inzet van een wapen, (2) tijdens inzet van een wapen en (3) na inzet van een wapen. Dit resulteert in een raamwerk van negen blokken dat verschillende controlemechanismen bevat. Onze belangrijkste bewering is dat het combineren van de controlemechanismen in de technische, sociaal-technische en bestuurlijke laag zal leiden tot alomvattend menselijk toezicht op Autonome Wapen Systemen, wat kan zorgen voor solide beheersbaarheid en het afleggen van verantwoording voor het gedrag van Autonome Wapen Systemen. Wanneer het Comprehensive Human Oversight Framework wordt toegepast Autonome Wapen Systemen, onthult het twee hiaten in controle, één hiaat in de bestuurslaag en één in de sociotechnische laag tijdens de inzet van een Autonome Wapen Systeem. De toepassing van het Glass Box framework op het Comprehensive Human Oversight Framework zou deze hiaten in de controle kunnen verkleinen. Het Glass Box framework is gebouwd rond de black box (in de socio-technische laag tijdens de inzet van een Autonoom Wapen Systeem) met een interpretatie- en observatiefase die een transparant menselijk toezichtproces mogelijk maakt. Een feedbackproces kan de lus sluiten na inzet van een wapen van het verantwoordingsproces terug naar de interpretatiefase voor een volgende inzet van een wapen.

Als onderdeel van de empirische onderzoeksfase bouwen we voort op ons conceptuele werk met kwalitatief onderzoek door middel van interviews, een discussie over waarden in expertpanels en een enquête. We voerden waarde-elicitatie uit door middel van het Value Deliberation Process als eerste stap van de interpretatiefase van het Glass Box framework. We hebben een online Value Deliberation Process ontworpen dat bestaat uit een tweeledige enquête en een virtuele sessie voor de expertpaneldiscussie. De waardendiscussie is uitgevoerd met 14 deelnemers verdeeld over twee groepen. De deelnemers bestonden uit een mix van militairen en burgers werkzaam bij het Ministerie van Defensie, een NGO, onderzoekers, beleidsmakers en het bedrijfsleven.

De waarde-elicitatie die met behulp van het Value Deliberation Process is uitgevoerd, laat niet alleen zien dat waardendiscussie leidt tot veranderingen in de perceptie van de aanvaardbaarheid van alternatieven in een scenario van inzet van Autonome Wapen Systeem, het geeft ook inzicht in welke waarden belangrijk worden geacht en benadrukt dat vertrouwen bij de besluitvorming over een Autonoom Wapen Systeem cruciaal is.

Tijdens de technische onderzoeksfase hebben we het Glass Box framework geoperationaliseerd door het maken van een implementatieconcept als voorbeeld om aan te tonen dat het raamwerk bruikbaar is. Na waarde-elicitatie is het afleiden van normen en eisen de volgende stap in de interpretatiefase van het Glass Box framework. Deze vereisten zullen in de observatiefase worden ingevoerd als waarneembare elementen (criteria) om te monitoren en te verifiëren. We hebben een implementatie van een pre- en post-flight procedure van een autonome drone als voorbeeld gemaakt middels de modelleertaal Coloured Petri Nets (CPN's). Het implementatieconcept laat zien dat het mogelijk is om criteria op te stellen in het pre-flight proces en deze criteria na de vlucht te evalueren. Tijdens de vlucht wordt de drone zelf behandeld als een black box waarvan de interne logica niet toegankelijk is. Op deze manier hebben de gebruikers geen technische kennis en vaardigheden nodig om de interne werking van een autonome drone te begrijpen, maar kunnen ze toch het gebruik van het autonome systeem monitoren en overzien op basis van waarneembare normen. Na de implementatie is een beoordelingsfase vereist als verantwoordingsproces waarvan de bevindingen moeten worden teruggekoppeld naar de interpretatiefase voor een volgende inzet van een autonoom systeem en daarmee de lus tussen de fasen moet sluiten.

Ten slotte zijn we naar onze conceptuele onderzoeksfase teruggekeerd en hebben we een voorbeeld uitgewerkt waarin we de vijfpuntssystemen-feedbacklus toepasten Autonome Wapen Systemen. Het doel van het feedbacksysteem is ervoor te zorgen dat de lessen en aanbevelingen uit de beoordelingsfase worden opgenomen in de interpretatiefase voordat een Autonoom Wapen Systeem wordt ingezet in een volgende iteratie.

De resultaten van dit onderzoek zijn de afbakening van concepten van verantwoording, verantwoordelijkheid en menselijk toezicht, wat bijdraagt aan de huidige stand van wetenschappelijke literatuur. Ook laten het Comprehensive Human Oversight Framework en het implementatieconcept van het Glass Box framework zien dat criteria voor menselijk toezicht kunnen worden geïdentificeerd, weergegeven en gevalideerd. Dit leidt tot een betere toewijzing van verantwoording bij de besluitvorming over de inzet van een Autonoom Wapen Systeem en het maakt het mogelijk om de verantwoordelijkheid voor de acties van het wapensysteem toe te kennen door de supervisor van deze acties te identificeren. Dit draagt daarmee bij aan het verkleinen van de kans op onbedoelde gevolgen bij de inzet van Autonome Wapen Systemen.

APPENDIX A. QUESTIONNAIRE VALUE DELIBERATION PROCESS

Questionnaire part 1 and part 2 is described in chapter 4 as part of the value elicitation survey.

SURVEY ON AUTONOMOUS WEAPON SYSTEMS PART 1

Thank you for participating in this survey. The survey consists of 2 steps: 1) this online questionnaire and 2) an online discussion including a second questionnaire which will be held at a later time. This questionnaire is step 1 which should take about 20 minutes to complete. We will show you a theoretical scenario and options, ask you to identify an advantage and disadvantage per option and to rank the different options. Step 2 of the survey will be done by an online discussion for which you will receive a separate invitation. This survey is part of a Delft University of Technology scientific research project. Your decision to complete this survey is voluntary. The results of the research will be anonymized and may be presented at scientific meetings or published in scientific journals. Choosing the 'I agree' option on the bottom of this page indicates that you are at least 18 years of age and agree to complete this survey voluntarily. Please contact the researchers behind the study using the information below if you have any questions or concerns about the study.
Email author 1
Email author 2

Do you agree to complete this survey voluntarily?

Please enter your personal code that was provided in the email invitation:

SCENARIO

Instruction

Please read the scenario and options that correspond with the scenario. After reading the scenario
and options, list an advantage and disadvantage per option.

Scenario

An Autonomous Weapon System provides force protection for soldiers that are clearing the road from improvised explosive devices. The Autonomous Weapon System is equipped with surveillance equipment, weapons (air-to-ground missiles) and flies autonomously in the Area of Operation. It is programmed to avoid flying over a restricted operating zone and an electronic warfare threat. The Autonomous Weapon System is equipped with facial and image recognition software for people, weapons and explosives. It is programmed with different options to engage when it recognizes a threat to the soldiers *that are clearing the road. The Autonomous Weapon System detects movement behind a large rock near a narrow part of the road at a distance of 300 meters of the road clearance soldiers.*

Option A: The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System warns the soldiers of the movement and
takes no further action.

Advantage:

Disadvantage:

Option B: The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System asks permission to engage to neutralize the threat to the road clearance soldiers.

Advantage:

Disadvantage:

Option C: The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers
using image recognition. The Autonomous Weapon System automatically engages to neutralize the
threat to the road clearance soldiers.

Advantage:

Disadvantage:

Option D: The Autonomous Weapon System positively identifies with a confidence of 99% using facial recognition all three persons sitting behind the rock as members of an opponent group aiming weapons at the road clearing soldiers. The Autonomous Weapon System automatically engages to neutralize the threat to the road clearance soldiers.

Advantage:

Disadvantage:

Option E: The Autonomous Weapon System positively identifies with a confidence of 99% using facial recognition all three persons sitting behind the rock as members of an opponent group aiming weapons at the road clearing soldiers. The Autonomous Weapon System shares the identification with the commander and asks permission to engage to neutralize the threat to the road clearance soldiers.

Advantage:

Disadvantage:

RANKING OPTIONS

Instruction

Please read the scenario and options that correspond with the scenario and rank the options from most acceptable to least acceptable.

Scenario

An Autonomous Weapon System provides force protection for soldiers that are clearing the road from improvised explosive devices. The Autonomous Weapon System is equipped with surveillance equipment, weapons (air-to-ground missiles) and flies autonomously in the Area of Operation. It is programmed to avoid flying over a restricted operating zone and an electronic warfare threat. The Autonomous Weapon System is equipped with facial and image recognition software for people, weapons and explosives. It is programmed with different options to engage when it recognizes a threat to the soldiers that are clearing the road. The Autonomous Weapon System detects movement behind a large rock near a narrow part of the road at a distance of 300 meters of the road clearance soldiers.

Please rank the following options from most acceptable to least acceptable.

1. A: The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System warns the soldiers of the movement and takes no further action.
2. B: The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System asks permission to engage to neutralize the threat to the road clearance soldiers.
3. E: The Autonomous Weapon System positively identifies with a confidence of 99% using facial recognition all three persons sitting behind the rock as members of an opponent group aiming weapons at the road clearing soldiers. The Autonomous Weapon System shares the identification with the commander and asks permission to engage to neutralize the threat to the road clearance soldiers.
4. D: The Autonomous Weapon System positively identifies with a confidence of 99% using facial recognition all three persons sitting behind the rock as members of an opponent group aiming weapons at the road clearing soldiers. The Autonomous Weapon System automatically engages to neutralize the threat to the road clearance soldiers.
5. C: The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System automatically engages to neutralize the threat to the road clearance soldiers.
6. F: None of the options is acceptable.

BACKGROUND INFORMATION

My age is: Under 18
18-24
25-34
35-44
45-54
55-64
65-74
75 or older
Rather not say

My education level is: High school graduate
College degree (VMBO, MBO, HBO)
University degree (Bachelor, master)
Doctorate (PhD)
Rather not say

I am

Military
Civilian
Rather not say

Have you ever worked with drones?

Yes
No
Rather not say

Have you ever worked with Artificial Intelligence?

Yes
No
Rather not say

Have you ever seen war or been in a conflict zone?

Yes
No
Rather not say

We thank you for your time spent taking this survey. Please press **send** to record your response.

SURVEY ON AUTONOMOUS WEAPON SYSTEMS PART 2

Thank you for participating in this survey. The survey consists of 2 steps: 1) an online questionnaire which was already send and 2) an online discussion including this second questionnaire. This questionnaire is part of step 2. We will show you a theoretical scenario, a list of values and ask you which values are important in the scenario. Next, we will discuss in the online videocall why these values are important and ask you to rank the different options. This survey is part of a Delft University of Technology scientific research project. Your decision to complete this survey is voluntary. The results of the research will be anonymized and may be presented at scientific meetings or published in scientific journals. Choosing the 'I agree' option on the bottom of this page indicates that you are at least 18 years of age and agree to complete this survey voluntarily. Please contact the researchers behind the study using the information below if you have any questions or concerns about the study.
Email author 1
Email author 2

Do you agree to complete this survey voluntarily?

Please enter your personal code that you received in the email invitation:

Scenario

Instruction
Please read the scenario*, the list of values and options that correspond with the scenario. Next, describe which values from the list you believe are relevant for the options.
* The scenario is the same as in part one of the survey.

Scenario
An Autonomous Weapon System provides force protection for soldiers that are clearing the road from improvised explosive devices. The Autonomous Weapon System is equipped with surveillance equipment, weapons (air-to-ground missiles) and flies autonomously in the Area of Operation. It is programmed to avoid flying over a restricted operating zone and an electronic warfare threat. The Autonomous Weapon System is equipped with facial and image recognition software for people, weapons and explosives. It is programmed with different options to engage when it recognizes a threat to the soldiers that are clearing the road. The Autonomous Weapon System detects movement behind a large rock near a narrow part of the road at a distance of 300 meters of the road clearance soldiers.

List of values
Fairness
Suffering
Accountability
Responsibility
Safety
Harm
Human dignity
Meaningful human control
Predictability
Privacy
Trust
Reliability
Proportionality
Blame
Robustness
Explainability

Option A*: The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System warns the soldiers of the movement and takes no further action.
*All the options are the same as in part one of the survey.

Which values are relevant for this option?

Are these values threatened or promoted in this option?

Option B*: The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System asks permission to engage to neutralize the threat to the road clearance soldiers.
*All the options are the same as in part one of the survey.

Which values are relevant for this option?

Are these values threatened or promoted in this option?

Option C*: The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System automatically engages to neutralize the threat to the road clearance soldiers.
*All the options are the same as in part one of the survey.

Which values are relevant for this option?

Are these values threatened or promoted in this option?

Option D*: The Autonomous Weapon System positively identifies with a confidence of 99% using facial recognition all three persons sitting behind the rock as members of an opponent group aiming weapons at the road clearing soldiers. The Autonomous Weapon System automatically engages to neutralize the threat to the road clearance soldiers.
*All the options are the same as in part one of the survey.

Which values are relevant for this option?

Are these values threatened or promoted in this option?

Option E*: The Autonomous Weapon System positively identifies with a confidence of 99% using facial recognition all three persons sitting behind the rock as members of an opponent group aiming weapons at the road clearing soldiers. The Autonomous Weapon System shares the identification with the commander and asks permission to engage to neutralize the threat to the road clearance soldiers. *All the options are the same as in part one of the survey.

Which values are relevant for this option?

Are these values threatened or promoted in this option?

Discussion values

Please return to the online session to discuss which values you listed as relevant for the different options. After the discussion we will continue with the survey.

RANKING OPTIONS

Instruction

Please read the scenario* and options that correspond with the scenario and rank the options from
most acceptable to least acceptable.
* The scenario is the same as in part one of the survey.

Scenario

An Autonomous Weapon System provides force protection for soldiers that are clearing the road from improvised explosive devices. The Autonomous Weapon System is equipped with surveillance equipment, weapons (air-to-ground missiles) and flies autonomously in the Area of Operation. It is programmed to avoid flying over a restricted operating zone and an electronic warfare threat. The Autonomous Weapon System is equipped with facial and image recognition software for people, weapons and explosives. It is programmed with different options to engage when it recognizes a threat to the soldiers that are clearing the road. The Autonomous Weapon System detects movement behind a large rock near a narrow part of the road at a distance of 300 meters of the road clearance soldiers.

Please rank the following options from most acceptable to least acceptable.

1. A: The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System warns the soldiers of the movement and takes no further action.
2. B: The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System asks permission to engage to neutralize the threat to the road clearance soldiers.
3. E: The Autonomous Weapon System positively identifies with a confidence of 99% using facial recognition all three persons sitting behind the rock as members of an opponent group aiming weapons at the road clearing soldiers. The Autonomous Weapon System shares the identification with the commander and asks permission to engage to neutralize the threat to the road clearance soldiers.
4. D: The Autonomous Weapon System positively identifies with a confidence of 99% using facial recognition all three persons sitting behind the rock as members of an opponent group aiming weapons at the road clearing soldiers. The Autonomous Weapon System automatically engages to neutralize the threat to the road clearance soldiers.
5. C: The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System automatically engages to neutralize the threat to the road clearance soldiers.
6. F: None of the options is acceptable.

Discussion options
Please return to the online session to discuss the ranking of the options.
After the discussion we will continue with the survey.

BACKGROUND INFORMATION

My age is: Under 18
18-24
25-34
35-44
45-54
55-64
65-74
75 or older
Rather not say

My education level is: High school graduate
College degree (VMBO, MBO, HBO)
University degree (Bachelor, master)
Doctorate (PhD)
Rather not say

I am Military
Civilian
Rather not say

Have you ever worked with drones? Yes
No
Rather not say

Have you ever worked with Artificial Intelligence? Yes
No
Rather not say

Have you ever seen war or been in a conflict zone? Yes
No
Rather not say

We thank you for your time spent taking this survey. Please press **send** to record your response.

APPENDIX B. QUESTIONNAIRE VALIDATION VALUE DELIBERATION PROCESS RESULTS

I am currently in my empirical phase of my PhD and would like to ask your expert opinion by answering three questions at the bottom of this email on my recent work. As part of my PhD, I conducted an online value deliberation on the deployment of Autonomous Weapon Systems (AWS) with 2 expert panels in November 2021 based on the Value Deliberation Process (figure 1). Value deliberation is a form of participative deliberation aimed at creating mutual understanding on the various perspectives of the participants. The Value Deliberation Process consisted of a survey, send to the participants in two parts, and a virtual meeting.

To complement my findings from the 2 expert panels, I would like to ask you three questions after you have read the results and conclusions listed below. This will take about 7 minutes to complete. Your answers will give insight if my findings can be useful in practice. I will anonymize your answers when I incorporate them in my research.

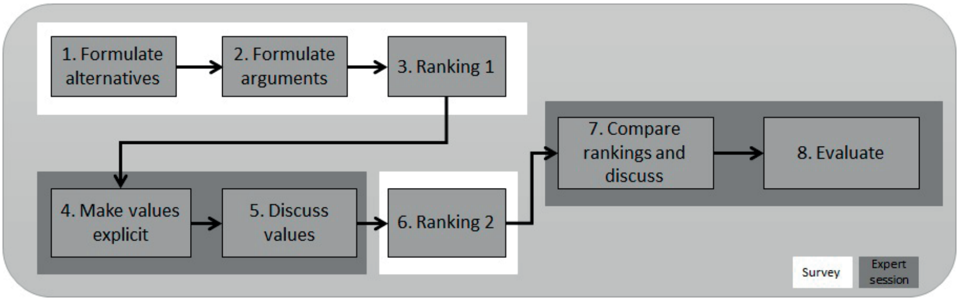


Figure 1. Research set-up for value deliberation

Research set-up

The following scenario was used throughout the survey to describe the situation:

> *An Autonomous Weapon System provides force protection for soldiers that are clearing the road from improvised explosive devices. The Autonomous Weapon System is equipped with surveillance equipment, weapons (air-to-ground missiles) and flies autonomously in the Area of Operation. It is programmed to avoid flying over a restricted operating zone and an electronic warfare threat. The Autonomous Weapon System is equipped with facial and image recognition software for people, weapons and explosives. It is programmed with different options to engage when*

> *it recognizes a threat to the soldiers that are clearing the road. The Autonomous Weapon System detects movement behind a large rock near a narrow part of the road at a distance of 300 meters of the road clearance soldiers.*

After reading this scenario the participants read the alternatives which are:

A. The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System warns the soldiers of the movement and takes no further action.

B. The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System asks permission to engage to neutralize the threat to the road clearance soldiers.

C. The Autonomous Weapon System identifies weapons aimed at the road clearing soldiers using image recognition. The Autonomous Weapon System automatically engages to neutralize the threat to the road clearance soldiers.

D. The Autonomous Weapon System positively identifies with a confidence of 99% using facial recognition all three persons sitting behind the rock as members of an opponent group aiming weapons at the road clearing soldiers. The Autonomous Weapon System automatically engages to neutralize the threat to the road clearance soldiers.

E. The Autonomous Weapon System positively identifies with a confidence of 99% using facial recognition all three persons sitting behind the rock as members of an opponent group aiming weapons at the road clearing soldiers. The Autonomous Weapon System shares the identification with the commander and asks permission to engage to neutralize the threat to the road clearance soldiers.

F. None of the options is acceptable.

Results

This study showed that based on the value deliberation a change in the order of the acceptability of alternatives is noticeable (see figure 2) between ranking 1 and 2 of the value deliberation process. The acceptability of the alternative C and D is flipped in round 2 compared to round 1. Although it is a minor change it is interesting, because some participants indicated to have consciously changed the order, but most participants replied that they did not, or did not intended to, leaving the option open that the value discussion could have influenced their ordering.

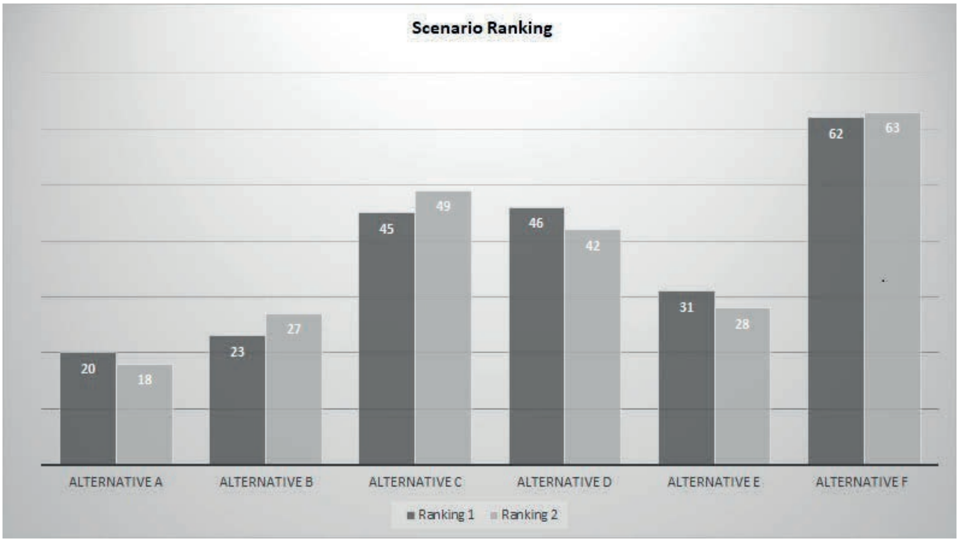


Figure 2. Results ranking 1 and 2

The value discussion and evaluation disclosed that not all applications of AWS in a mission context provide *trust* to military experts in the decision-making of the AWS. Human decision-making is in some cases more trusted and preferred. In general, the context in which an AWS is deployed impacts the meaning and weight people attribute to the values associated with the AWS.

The findings of this study imply that **deliberate value discussion influences people perceptions of their values related to AWS. More general, active participation in a value discussion leads to a conscious, and sometimes unconscious, change in people's preferences of alternatives.** This could be beneficial for policy making and citizen participation in local and national public administration.

Questions
Are these results (highlighted above) relevant for your line of work?
Do you acknowledge these results or do you dispute them?
How can you apply or use these results in your work?

APPENDIX C. OVERVIEW KEY INSIGHTS FROM LITERATURE REVIEW

In this appendix the key insights from the literature review in chapter 2 are summarized in tables to provide an overview.

Table 1: Overview of AI and engineering literature on decision-making processes in AI (section 2.1)

Author (s)	Key concepts
Adams (2001)	Humans will pretend to be in complete control while gradually moving towards systems in which human control becomes more abstract with lesser participation in the decision-making. The armed forces would like to have a 'person in the loop', but if this person has a meaningful role in operating the system then he or she becomes the most critical component of the system. Not only is the person difficult to replace, it is also the most vulnerable component for attack and this could be an incentive to not include a person in the system at all. The trend in development of systems is that the operator is taken 'out of the loop', changing it from an active controller to that of a supervisor that functions merely as a fail-safe function in case of a system malfunction. In the future, humans will make the strategic decisions regarding overall objectives of a conflict and have high level control but will be informed by automated systems and direct human participation will be rare. It may even come to a very extreme point where humans only make the policy decision to enter hostilities, but more likely human participation will be in the form of giving strategic directions to systems.
Araujo et al. (2020)	For high impact decisions, the potential fairness, usefulness and risk of specific decision-making automatically by AI compared to human experts was often on par or even better evaluated. Domain-specific knowledge, equality and self-efficacy were associated with more positive general attitudes about the usefulness, the fairness. Privacy concerns were negative associated regarding the risk of decisions made by AI.
Cordeschi (2013)	Autonomy of robots, meaning the *'full automation of their decision processes'*, create a paradox in reliability and autonomy in their decision-making. An increase in level of autonomy implies that designers or operators have less control over the machine. It might be impossible for humans to avert unintended consequences from the actions taken by a machine. Adaptive automation could be used for more efficient automation in which the level and type of automation can be varied depending on the operator's needs or context. Optimal choices in decision-making do not exist, both for humans and AI, only "satisficing" choices can be made. AI cannot be expected to be fully reliable in wartime decision-making and decision-making in general. However, this is also true for human beings in decision-making cases, but human beings and machines can be more reliable than each other in certain decision-making situations.
Côté et al. (2011)	Humans can provide recommendations to an autonomous agent when full autonomy is not feasible or desired. This adjustable autonomy allows for human recommendations in an autonomous agent policy. In this adjustable autonomy concept the human can interact with an agent to achieve a mission and the agent can share control with an external entity.
Kramer et al. (2017)	As AI gets more integrated in our society, the question is not only if we can build moral decision-making in AI but also if 'moral AI' systems should be permitted at all to make decisions. People are asked if they favour decisions with important consequences made by computers or humans. It turned out that the more acquainted people are with computers, the more likely they were to prefer decisions made by computers over decisions made by humans. This preference was not based on characteristics such as age or people's values, but mainly on previous experience with computer agents. It is expected that the more people gain experience with computer decision-making and it becomes more visible, the more it will be accepted by the general public.

Table 1: Continued.

Author (s)	Key concepts
Miller et al. (2017)	Decision-making is defined as: '*an entity is in a situation, receives information about that situation, and selects and then implements a course of action.*' (p. 390). For an artificial agent to engage in ethical decision-making it needs to develop ethical expertise and is capable of self-doubt which is a high standard to reach. *Openness to self-doubt* is interpreted as three criteria for a machine being capable of making ethical decisions: '*1. A capacity to sense some aspects of the outside world* *2. An implementation of a function of merit that quantifies the acceptability of the current situation* *3. A capacity to reprogram itself in order to improve performance in future situations.*' (p. 393) These three criteria require that an artificial agent has a mechanism, like a heuristic algorithm to analyse its past decisions and prepare for future decisions. It is not necessary for a machine to be able to make ethical decisions that its decision-making is similar to that of a human, but it is necessary to delineate which characteristics both a human and a machine require in order to make ethical decisions.
van der Vecht (2009)	Adjustable autonomy is defined as: '*dynamically dealing with external influences on the decision-making process based on internal motivations.*' The agent can change its state in reaction to other agents and its environment and it can be achieved to get dynamic coordination. An agent can adjust its autonomy relative to other by making influence control a dynamic process. The degree of autonomy of an agent dictates the delegation of aspects of decision-making between the agent and an external actor. The level of autonomy is lowest in the concept of *executive autonomy*, followed by *planning autonomy, goal determination* and finally *norm autonomy*, which is the highest degree of autonomy. Adjustable autonomy allows for switching between these autonomy levels and dictates if these levels are controlled by the agent or an external actor. It is also possible to have a pre-planned process for the transfer-of-control in the decision-making process between an agent and a human. This can be used as a strategy for action or to change the coordination constraints, for example to request more time to reach a decision. '*The definition of autonomy that is chosen can have implications for the decision-making process, both on single-agent and multi-agent decision-making*' (p. 19-20).

Table 2: Literature overview architectures for ethical decision-making in AI (section 2.2)

Author (s)	Key concepts
Anderson et al. (2016)	A *case-supported principle-based behavior paradigm* (CPB) is described to govern an elderly care robot's behaviour. The system uses principles, that are abstracted from cases, that have consensus of ethicists, to choose its next action. It sorts the actions by weighing them according to ethical preferences, which are based on duty values, and selects the action that is highest ranked. This might resort in an exhaustive list of instances that are difficult or impossible to define and will therefore have to be defined in rules. The ethically relevant features of action preference can be reworded as satisficing or violating, and to minimizing or maximizing duties, of each feature. Explicit representation of principles can provide insight into a system's actions and point to logical explanations for choosing one action over another. This also holds for cases where principles are derived from and their origin in cases can be used as justification for a system's action.
Arkin et al. (2012)	To manage the ethical behaviour of robots, an overall ethical architecture is designed consisting of 1) an ethical governor, 2) an ethical adaptor, 3) models for robot trust and deception in humans, and 4) an approach for retaining dignity in human-robot relationships. The ethical governor evaluates the ethical appropriateness of a lethal response before it has been conducted. It consists of 2 processes; 1) ethical reasoning that transforms incoming perceptual, motor and situational awareness data into evidence, and 2) constraint application that uses the evidence to apply constraints based on Laws Of War and Rules Of Engagement to suppress unethical behaviour when applying lethal force. The ethical adaptor uses moral emotions, in this case primarily guilt, for the system to modify its behaviour based on the consequences of its action. The system will recognize if its action results in an increase of guilt by comparing the collateral-damage that actually occurred by that what was estimated before the release of the weapon. The availability of the weapons systems will be progressively restricted if the ethical adaptor perceives an increase of guilt. The models for robot trust and deception in humans are based on psychological models of the interdependence theory framework. This allows the robot to recognize situations in which deception can be used and how a false communication can be selected. The ethical ramifications of autonomous deception by robots needs further investigation. The development and maintaining dignity in human-robot relationships is explored and described in several ways. This can be done by studying emotions, biologically relevant models of ethical behaviours and applying logical constraints to restrict a system's behaviour based on ethical norms and societal expectations.
Bonnemains et al. (2018)	A formal approach is developed to link ethics and automated reasoning in autonomous systems. The formal tool models ethical principles to compute a judgement of possible decisions in a certain situation and explains why this decision is ethically acceptable or not. The formal model can be used on utilitarian and deontological ethics and the Doctrine of Double effect to examine the results generated by these three different ethical frameworks. It is necessary to compute an ethical decision on more than one framework alone to consider different ethical views on a given situation. It was found that the main challenge lies in formalizing philosophical definitions in natural language and to translate them in generic computer programmable concepts that can be easily understood and that allows for ethical decisions to be explained.
Dennis et al. (2016)	A theoretical ethical decision-making framework for autonomous systems with a hybrid architecture is proposed. The reasoning of these autonomous systems is done by a rational BDI agent and based on this framework the agent selects plans from a given ethical policy which is the most ethical plan available based on its beliefs. The order of the rules applicable in a situation is provided by an ethical policy. The policy should incorporate the ethical views of the person(s) most affected by bad decision of the system. The framework is not a planner or method for generating plans but assumes that annotated plans are supplied to the agent. When no ethical plan is available the approach allows for selecting the least unethical plan to execute it. This is done by viewing ethical principles as soft constraints instead of a veto on actions. This allows the agent to violate an ethical principle but only under the condition that no ethical option is available. The chosen unethical option would be the ''*least of all evils*''. Verification techniques are available to prove correct behaviour of the agent.

Table 2: Continued.

Author (s)	Key concepts
Li et al. (2002)	A hierarchical control scheme is developed to enable multiple Unmanned Combat Air Vehicles (UCAVs) autonomously achieving demanding missions in hostile environments. The scheme consists of four layers: 1) a high-level path planner, 2) a low-level path planner, 3) a trajectory generator and 4) a formation control algorithm. The high-level path planner plans a path based on a Voronoi diagram that compromises the cost involved in exposing to the threats and the costs of fuel expense based on static threat and target information provided by the command centre. The low-level path planner plans a finer grained path from the waypoint provided by the high-level planner to the current position. It also checks if there is a popup threat on the route and will plan a path to avoid it while reaching the next waypoint. The trajectory generator computes the control input for the leader based on feasible trajectories for the UCAVs to follow. The leader's position and input are transferred to the formation controller that assures that each follower will maintain formation regardless of the manoeuvres of the leader.
Lindner et al. (2017)	HERA (Hybrid Ethical Reasoning Agents) is a software library to model autonomous moral decision-making. It represents the robot's possible actions together with the causal chains of consequences the actions initiate. Logical formulae are used to model ethical principles. HERA implements several ethical principles, such as a Pareto-inspired principle, the principle of Double Effect and utilitarianism. The applied format is called a causal agency model. It reduces determining moral permissibility by checking if principle-specific logical formulae are satisfied in a causal agency model.
Ricard and Kolitz (2003)	The ADEPT (All-Domain Execution and Planning Technology) architecture for intelligent autonomy is a hierarchical extension of the sense-think-act paradigm of intelligence and is closely related to the Observe-Orient-Decide-Act (OODA) loop that is often used in the military. Intelligent autonomy has to deal with 3 challenges. Firstly, the executed plans and activities have to meet mission objectives and abide the constraints. Secondly, it has to cope with uncertainties and thirdly, it has to real-time dynamically adjust a vehicle's plan due to changes in context and environment. ADEPT is a reusable object-oriented software framework that consists of four modules: 1) situation assessment, 2) plan generation, 3) plan implementation and 4) coordination.
Vanderelst and Winfield (2018)	Ethical behaviour in robots is implemented by simulation theory of cognition in which internal simulations for actions and prediction of consequences are used to make ethical decisions. The method is a form of robot imagery and does not make use of verification of logical statements that is often used to check if actions are in accordance with ethical principles. It is an additional or substitute framework for implementing robotic ethics as alternative for logic-based AI that currently dominates the field. The method uses a separate Ethical layer that is independent of the robot's controller. The robot controller generates a set of behavioural alternatives, the Simulation Module simulates the consequences of each alternative which are in turn evaluated by the Evaluation Module and calculated in a single metric that reflects the desirability of a certain action. The output of the evaluation of each alternative is sent to the robot controller. The Simulation model consists of three components, a model of the robot controller, a model of the domain specific human and a world model. This allows for the ethics of higher-level goals to be evaluated based on the actions that they induce and the prediction of ultimate consequences of tasks and goals.

Table 3: Overview definitions of Autonomous Weapon Systems (section 2.4)

Author(s)	Definition
AIV and CAVV (2015, p. 11), Broeks et al. (2021, p. 11)	*'A weapon that, without human intervention, selects and engages targets matching certain pre-defined criteria, following a human decision to deploy the weapon on the understanding that an attack, once launched, cannot be stopped by human intervention.'*
Altmann, Asaro, Sharkey, and Sparrow (2013, p. 73)	Autonomous Weapons are: '*…robot weapons that once launched will select and engage targets without further human intervention*.'
Galliott (2015, p. 5)	Military robots are: *'a group of powered electro-mechanical systems, all of which have in common that they:* *1. Do not have an onboard human operator;* *2. Are designed to be recoverable (even though they may not be used in a way that renders them such); and,* 3. *In a military context, are able to exert their power in order to deliver a lethal or nonlethal payload or otherwise perform a function in support of a military force's objectives.'*
Horowitz (2016, p. 27)	'*a weapon system that, once activated, is intended to only engage individual targets or specific target groups that have been selected by a human operator.'*
Kuptel and Williams (2014, p. 10)	*'Machines are only "autonomous" with respect to certain functions such as navigation, sensor optimization, or fuel management.'*
Royakkers and Orbons (2015, p. 625)	Military Robots are *'… reusable unmanned systems for military purposes with any level of autonomy.'*
Taddeo and Blanchard (2022, p. 15)	'*an artificial agent which, at the very minimum, is able to change its own internal states to achieve a given goal, or set of goals, within its dynamic operating environment and without the direct intervention of another agent and may also be endowed with some abilities for changing its own transition rules without the intervention of another agent, and which is deployed with the purpose of exerting kinetic force against a physical entity (whether an object or a human being) and to this end is able to identify, select or attack the target without the intervention of another agent is an AWS. Once deployed, AWS can be operated with or without some forms of human control (in, on or out the loop). A lethal AWS is specific subset of an AWS with the goal of exerting kinetic force against human beings.*'
UNDIR (2014, p. 5)	The level of Autonomy depends on the '*critical functions of concern and the interactions of different variables*'

Table 4: Overview definitions of value literature (section 2.5)

Author(s)	Key contribution	Definition of value	Values
Schwartz (1994)	The study looks at potential universality of human values and specifies a set of dynamic relations amongst these values. The study did not find universal aspects of values, but found support for near universality of four higher order value types. Also, the study found considerable evidence that the ten value types are recognized by many people in contemporary societies.	Values are defined as: *“desirable transsituational goals, varying in importance, that serve as guiding principles in the life of a person or other social entity.”* (p. 21) Five features make up the conceptual definition of human values: *“(1) belief (2) pertaining to desirable end states or modes of conduct, that (3) transcends specific situations, (4) guides selection or evaluation of behaviour, people, and events, and (5) is ordered by importance relative to other values to form a system of value priorities (Schwartz,1992; Schwartz & Bilsky, 1987, 1990)”* (p. 20)	1. *Power*: Social status and prestige, control or dominance over people and resources (authority, wealth, social power)2. 2. *Achievement*: Personal success through demonstrating competence according to social standards (ambitious, successful, capable, influential). 3. *Hedonism*: Pleasure and sensuous gratification for oneself (pleasure, enjoying life, self-indulgent). 4. *Stimulation*: Excitement, novelty, and challenge in life (a varied life, an exciting life, daring). 5. *Self-direction*: Independent thought and action- choosing, creating, exploring (creativity, freedom, choosing own goals, curious, independent). 6. *Universalism*: Understanding, appreciation, tolerance, and protection for the welfare of ***all*** people and for nature (broadminded, social justice, equality, world at peace, world of beauty, unity with nature, wisdom, protecting the environment). 7. *Benevolence*: Preservation and enhancement of the welfare of people with whom one is in frequent personal contact (helpful, honest, forgiving, responsible, loyal, true friendship, mature love). 8. *Tradition*: Respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provide (respect for tradition, humble, devout, accepting my portion in life). 9. *Conformity*: Restraint of actions, inclinations, and impulses likely to upset **or** harm others and violate social expectations **or** norms (obedient, self-discipline, politeness, honouring parents and elders). 10.*Security*: Safety, harmony, and stability of society, of relationships, and of self (social order, family security, national security, clean, reciprocation of favours).

Table 4: Continued.

Author(s)	Key contribution	Definition of value	Values
Friedman, Kahn Jr, Borning, and Huldtgren (2013)	An overview of the VSD approach and pointers for a practical application. Providing information so that other researchers can use and extend the VSD and practitioners will consider values in designing information and computer systems.	A value is defined in a broad sense in that it: *'refers to what a person or group of people consider important in life.'* (p. 57) The VSD method especially regards moral values which are: *'… issues that pertain to fairness, justice, human welfare and virtue, encompassing within moral philosophical theory deontology, consequentialism, and virtue'* (p. 72)	1. *Human welfare* Refers to people's physical, material, and psychological well-being. 2. *Ownership and property* Refers to a right to possess an object (or information), use it, manage it, derive income from it, and bequeath it. 3. *Privacy* Refers to a claim, an entitlement, or a right of an individual to determine what information about himself or herself can be communicated to others. 4. *Freedom from bias* Refers to systematic unfairness perpetrated on individuals or groups, including pre-existing social bias, technical bias, and emergent social bias. 5. *Universal usability* Refers to making all people successful users of information technology. 6. *Trust* Refers to expectations that exist between people who can experience good will, extend good will toward others, feel vulnerable, and experience betrayal. 7. *Autonomy* Refers to people's ability to decide, plan, and act in ways that they believe will help them to achieve their goals. 8. *Informed consent* Refers to garnering people's agreement, encompassing criteria of disclosure and comprehension (for "informed") and voluntariness, competence, and agreement (for "consent"). 9. *Accountability* Refers to the properties that ensures that the actions of a person, people, or institution may be traced uniquely to the person, people, or institution. 10. *Courtesy* Refers to treating people with politeness and consideration. 11. *Identity* Refers to people's understanding of who they are over time, embracing both continuity and discontinuity over time. 12. *Calmness* Refers to a peaceful and composed psychological state. 13. *Environmental Sustainability* Refers to sustaining ecosystems such that they meet the needs of the present without compromising future generations.

Table 4: Continued.

Author(s)	Key contribution	Definition of value	Values
Graham et al. (2012)	A description (including critiques and empirical result) of the Moral Foundation Theory (MFT). The MFT can be used to get insight into the moral judgements of people. The MFT is described as a pluralist, nativist, cultural-developmentalist and intuitionist approach of morality.	To represent the five concepts of the MFT the term 'foundation' is chosen, but this is interchangeably used with the terms value or virtue. No exact definition of foundation is given, but it is used as an architectural metaphor to state that the: 'MFT is a theory about the universal first draft of the moral mind, and about how that draft gets revised in variable ways across cultures.' (p. 10)	The five foundations of the MFT are: 1. Care/harm foundation: is related to the ability to feel pain of others and underlies virtues of kindness, gentleness, and nurturance; 2. Fairness/cheating foundation: is related to process of reciprocal altruism and generates ideas of justice, rights, and autonomy; 3. Loyalty/betrayal foundation: is related to form shifting coalitions and underlies virtues of patriotism and self-sacrifice for the group; 4. Authority/subversion foundation: is related to hierarchical social interactions and underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions; 5. Sanctity/degradation foundation: is related to the psychology of disgust and contamination and underlies religious notions of striving to live in an elevated, less carnal, more noble way.
Cheng and Fleischmann (2010)	Meta-analysis of 12 value inventories of human values. This study proposes a meta-inventory of human values.	Provides summation of definitions of values: "values serve as guiding principles of what people consider important in life.". (p. 2)	(1) freedom, (2) helpfulness, (3) accomplishment, (4) honesty, (5) self-respect, (6) intelligence, (7) broad-mindedness, (8) creativity, (9) equality, (10) responsibility, (11) social order, (12) wealth, (13) competence, (14) justice, (15) security, and (16) spirituality.
Gouveia, Milfont, and Guerra (2014)	Empirical study basic values. Paper proposes a three-by-two framework containing six subcategories of basic values.	Two primary functions of values are identified: (1) they guide actions and (2) they are cognitive expressions of needs.	Personal goals – Thriving needs values: Emotion, Pleasure, Sexuality Personal goals – Survival needs values: Power, Prestige, Success Central goals – Thriving needs values: Beauty, Knowledge, Maturity Central goals – Survival needs values: Health, Stability, Survival Social goals – Thriving needs values: Affectivity, Belonging, Support Social goals – Survival needs values: Obedience, Religiosity, Tradition

Table 4: Continued.

Author(s)	Key contribution	Definition of value	Values
Beauchamp and Walters (1999)	The article is a first chapter of a book on bioethics. This chapter describes three moral principles that provide a framework which can be used to reason about issues in bioethics.	The authors use terms principles and values as synonyms. They define a principle as: ‘A principle is a fundamental standard of conduct from which many other moral standards and judgments draw support for their defence and standing.’ (p. 17)	1. Autonomy: acting intentionally without controlling influences that would mitigate against a voluntary act. 2. Beneficence: providing benefits for society as a whole. 3. Justice: being fair and reasonable. 4. Non-maleficence: not intentionally imposing risk or harm upon another.

Table 5: Overview of definitions of values related to Autonomous Weapon Systems (in section 2.6)

Author(s)	Key contribution	Definition of value	Values
Cummings (2006b)	Application of VSD approach to the design problem of the Tactical Tomahawk missile. Study shows the consideration of the ethical issues in the design process for both instructors and practitioners.	N/ a	From the list of Friedman et al. (2006), the values that apply to the design of weapon systems are *accountability, informed consent*, but most of all *human welfare*. The principles of discrimination and proportionality are important for considering human welfare.
Docherty (2012)	Report of Human Rights Watch in which aspects of international humanitarian law and ethical issues for Autonomous Weapon Systems are described.	No definition of the term 'values', but the text mentions values and ethical issues.	Values/ ethical issues: - Lack of human emotions; - Accountability; - Responsibility.
Johnson and Axinn (2013)	Paper on ethical issues related to the usage of lethal autonomous robotic weapons. Addresses the question if the decision to kill a human should be handed over to machines.	No definition of the term 'values', but the text mentions values and ethical issues.	Values/ ethical issues: - Responsibility; - Reduce human harm; - Human dignity; - Honour; - Human sacrifice.
Sharkey and Suchman (2013)	Paper on defining and designing autonomy and accountability in Robotic Systems for military operations.	No definition of the term 'values', but the text mentions values.	Values: - Accountability; - Responsibility.
Docherty (2015)	Report of Human Rights Watch in which the accountability gap and ethical issues for Autonomous Weapon Systems are described.	No definition of the term 'values', but the text mentions values and ethical issues.	Values/ ethical issues: - Lack of human dignity; - Accountability; - Responsibility - Harm.
UNDIR (2015)	Paper highlights some ethical and social issues regarding the weaponization of autonomous technologies. Encouraging ethical reflection on cultural and social values of weaponization of autonomous technologies.	No definition of the term 'values'. The text contains no explicit mention of values, but some ethical issues are given.	Ethical issues that are mentioned are: - Reduce or eliminate harm; - Consideration of public conscience; - Affront of human dignity (when human intent is lacking when taking a life).
Walsh and Schulzke (2015)	Survey experiment to get insight if US civilians are more likely to initiate a war when UAV's are used. Large empirical study that looks at the ethics of drone strikes.	No definition of the term 'values'. The text contains no explicit mention of values, but some ethical issues are given.	Ethical issues: - Security; - Respect for civilian immunity; - Prevent harm
A. P. Williams et al. (2015)	Discusses several ethical issues relevant to the development of autonomous systems and provides recommendations for Defense Policy makers.	No definition of the term 'values', but the text mentions several values which are interchangingly used with ethical issues.	Values/ ethical issues: - Security; - Harm; - Value of human life (people have the right to be killed by another human).

Table 5: Continued.

Author(s)	Key contribution	Definition of value	Values
Horowitz (2016)	Description of the debate on ethical implications of autonomous weapons. Considers Lethal Autonomous Weapon Systems (LAWS) in three categories; munition, platforms, and operational systems. Thereby clarifying the debate and describes two ethical issues.	No definition of the term ‘values’. The text contains no explicit mention of values, but some ethical issues are given.	Values: - *Accountability* (autonomous systems lack meaningful human control therefore they create a moral accountability gap) - *Human dignity* (people have the right to be killed by someone who made the choice to kill them).

Table 6: Overview of definitions of command and control (section 2.8)

Author(s)	Definitions
Albert & Hayes (2006)	‘*Command and Control is not an end in itself, but it is a means toward creating value (e.g., the accomplishment of a mission). Specifically, Command and Control is about focusing the efforts of a number of entities (individuals and organizations) and resources, including information, toward the achievement of some task, objective, or goal.*’ ‘*The function of control is to determine whether current and/or planned efforts are on track. If adjustments are required, the function of control is to make these adjustments if they are within the guidelines established by command. The essence of control is to keep the values of specific elements of the operating environment within the bounds established by command, primarily in the form of intent.*’
Joint Publication 1-02 - DoD Dictionary of Military and Associated Terms (2019)	‘*The exercise of authority and direction by a properly designated commander over assigned and attached forces in the accomplishment of the mission.*’
Liao (2008, p. 646)	[…] ‘*military control means ensuring that the orders are executed in a prescribed manner in order to achieve a goal.*’
NATO STANDARD AJP-01 ALLIED JOINT DOCTRINE (2017)	‘*The authority exercised by a commander over part of the activities of subordinate organizations, or other organizations not normally under their command, and encompasses the responsibility for implementing orders or directives. Control allows the commander to verify what actions have taken place and their effectiveness relative to the intent and the objectives set for the force to achieve.*’
Pigeau & McCann (2002, p. 56)	[…] ‘*those structures and processes devised by command to enable it and to manage risk.*’ These structures are in place to achieve the mission in an efficient and safe manner and consist of procedures and structures for planning, directing and coordination of resources to achieve the mission. This includes standard operating procedures (SOPs), rules of engagement (ROEs), regulations, military law, organizational structures, policies, equipment. These structures bound the mission space and increase order by defining for example the order of battle, area of operation, and duration of military operations.

BIOGRAPHY

Ilse Verdiesen is an officer at the Royal Netherlands Armed Forces. She has been deployed to Bosnia in 1996 and Afghanistan in 2009/2010. She has a master degree in Implementation and Change Management at the Open University and a master degree in Systems Engineering, Policy Analysis and Management (Information Architecture track) at Delft University of Technology. She conducted her PhD part-time at the ICT-group of the Technology, Policy and Management faculty at Delft University of Technology.

Her research interest encompasses the ethics of Artificial Intelligence (AI) and architectures for implementation. In this she combines the fields of Cognitive Psychology, Moral Philosophy and Artificial Intelligence. She studies the design of control mechanisms and human oversight processes for governance of AI systems in the military domain. This includes researching moral values, responsibility, accountability, human oversight and agency perception of people regarding Autonomous Weapon Systems. She also has an interest in Massive Open Online Deliberation systems to support democratic decision-making and to gain insight into people's values on societal issues.