Accountability and Control Over Autonomous Weapon Systems
A Framework for Comprehensive Human Oversight

Verdiesen, Ilse; Santoni de Sio, Filippo; Dignum, Virginia

DOI
10.1007/s11023-020-09532-9

Publication date
2020

Document Version
Final published version

Published in
Minds and Machines

Citation (APA)
Verdiesen, I., Santoni de Sio, F., & Dignum, V. (2020). Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight. *Minds and Machines*, *31*(1), 137-163. https://doi.org/10.1007/s11023-020-09532-9

Minds and Machines
https://doi.org/10.1007/s11023-020-09532-9

ORIGINAL ARTICLE

Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight

Ilse Verdiesen[1] · **Filippo Santoni de Sio**[1] · **Virginia Dignum**[1,2]

Received: 17 May 2020 / Accepted: 22 July 2020

Abstract
Accountability and responsibility are key concepts in the academic and societal debate on Autonomous Weapon Systems, but these notions are often used as high-level overarching constructs and are not operationalised to be useful in practice. "Meaningful Human Control" is often mentioned as a requirement for the deployment of Autonomous Weapon Systems, but a common definition of what this notion means in practice, and a clear understanding of its relation with responsibility and accountability is also lacking. In this paper, we present a definition of these concepts and describe the relations between accountability, responsibility, control and oversight in order to show how these notions are distinct but also connected. We focus on accountability as a particular form of responsibility—the obligation to explain one's action to a forum—and we present three ways in which the introduction of Autonomous Weapon Systems may create "accountability gaps". We propose a Framework for Comprehensive Human Oversight based on an engineering, socio-technical and governance perspective on control. Our main claim is that combining the control mechanisms at technical, socio-technical and governance levels will lead to comprehensive human oversight over Autonomous Weapon Systems which may ensure solid controllability and accountability for the behaviour of Autonomous Weapon Systems. Finally, we give an overview of the military control instruments that are currently used in the Netherlands and show the applicability of the comprehensive human oversight Framework to Autonomous Weapon Systems. Our analysis reveals two main gaps in the current control mechanisms as applied to Autonomous Weapon Systems. We have identified three first options as future work for the design of a control mechanism, one in the technological layer, one in the socio-technical layer and one the governance layer, in order to achieve comprehensive human oversight and ensure accountability over Autonomous Weapon Systems.

Keywords Autonomous Weapon Systems · Responsibility · Accountability · Accountability gap · Meaningful human control · Human oversight · Comprehensive human oversight framework

Extended author information available on the last page of the article

Published online: 01 August 2020

1 Introduction

Accountability and responsibility are key concepts in the academic and societal debate on the ethics and politics of Autonomous Weapon Systems. 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 (UN GGE LAWS 2018) lists '*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.*' as one of the possible guiding principles for LAWS. At the same time, many scholars express concerns that Autonomous Weapon Systems will lead to an "accountability gap" or "accountability vacuum", that is 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). According to the International Committee of the Red Cross (ICRC 2018), to uphold moral responsibility for decisions in the use of force, human control that is 'meaningful', 'effective' or 'appropriate' is needed and should be maintained. The notion of "Meaningful Human Control" is often mentioned as a requirement in the debate on Autonomous Weapon Systems, but a common definition of what this notion means in practice is lacking (Ekelhof 2019). Scholars have been working on defining and operationalizing Meaningful Human Control over the past years. Horowitz and Scharre (2015) were among the first to try to identify essential components for Meaningful Human Control and according to them Meaningful Human Control addresses accountability, moral responsibility and controllability. Roff and Moyes (2016) also characterized Meaningful Human Control based on three temporal layers; before the start of hostilities (ante-bellum), during the hostilities (in bello) and after the hostilities (post bellum). In their view, Meaningful Human Control links accountability systems and the need for responsible design—when the mechanisms in the first two layers fail, there is a need for accountability. However, both Horowitz and Scharre (2015) and Roff and Moyes (2016) do not sufficiently distinguish between the different forms of responsibility involved, mostly define control in terms of a relation between one human operator (or commander) and one technical device, and do not provide a clear definition of Meaningful Human Control.

The notions of accountability, moral responsibility and controllability are often used as high-level and overarching constructs in the Autonomous Weapon Systems debate, but their definition and relations are not always fully clarified. Nor is it clear how to operationalize these notions into working and verifiable requirements for practical use. In this article, we describe the relations between accountability, moral responsibility, controllability and oversight to show how these notions are distinct, but also connected in order to deepen our understanding. We purposely do not repeat the extensive philosophical discussion on Meaningful Human Control and accountability gaps, but aim to operationalize the

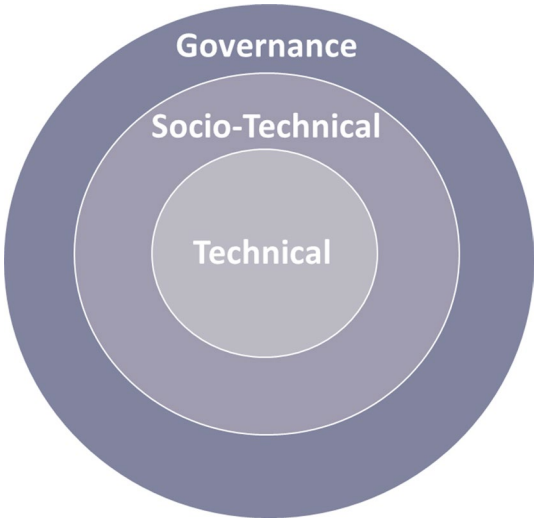

Fig. 1 Conceptualization cyberspace in layers (based on Van den Berg 2015)

concept of Human Oversight in a comprehensive approach and to provide concrete recommendations for an oversight process. We propose a framework for Comprehensive Human Oversight based on an engineering, socio-technical and governance perspective on control. By this, we broaden the view on the control over Autonomous Weapon Systems and take a comprehensive approach that may ensure solid controllability and accountability for the behaviour of 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 and provides an account of events after they occurred. A person needs to be accountable in order to be responsible and to have control to be able to account for his or her actions and an oversight mechanism is needed in order to hold an actor accountable (Caparini 2004; Fitzsimmons and Sangha 2010; Schedler 1999). Control is described in different ways in different disciplines. In the engineering perspective as mechanism to align goals and output of a technical system which is a mechanical or cybernetic view on control (Åström and Kumar 2014; Pigeau and McCann 2002). In the socio-technical perspective as the ability to induce the behaviour of another person (Koppell 2005) and is also applicable to entities and systems. Control can also be viewed from a governance perspective. Pesch (2015) noted that this perspective on control is lacking robust institutionalized frameworks for engineers.

We view Autonomous Weapon Systems as complex socio-technical systems which is best conceptualized as the result of the interaction between technical, socio-technical (human–machine) and governance components. Therefore, we take in this article the conceptual framework proposed by Van den Berg (2015) to characterize the cyber domain (cf. Fig. 1) and adapt it so as to describe the levels of controllability of Autonomous Weapon Systems. In his framework, Van den Berg (2015) describes the cyber domain constituted of three different layers: (1) the inner

layer is the *technological layer* in which the technology is described, (2) the middle layer is the *socio-technical layer* in which humans and technology interact in activities and (3) the outer layer is the *governance layer* in which institutions govern these activities.

In this article, we aim to operationalize the concepts of accountability, control and oversight from a design point-of-view and describe how these notions are both distinct and relate to each other. We realize that a rich and long philosophical discussion on these concepts exists. We focus on those aspects of these concepts that already described in literature, describe the relationships between them and apply them to the military domain. Our main claim is that combining the control mechanisms in the technical, socio-technical and governance layer will lead to a new concept of Comprehensive Human Oversight over Autonomous Weapon Systems which may ensure solid controllability and accountability for the behaviour of Autonomous Weapon Systems. By combining these three perspectives, we intend to broaden the current narrow view on Meaningful Human Control, which is often defined as the relationship between the human operator and Autonomous Weapon System.[1] By broadening this view to Comprehensive Human Oversight, we also consider the governance aspects concerning Autonomous Weapon Systems and in the second part of the paper we identify at least two gaps in the control mechanisms of Autonomous Weapon Systems.

The remainder of this article is as follows. Section 2 describes the relationship between responsibility and accountability by characterizing responsibility as being either *forward-looking*, *backward-looking* or *active* (Pesch 2015; Van de Poel 2011). We note that accountability is a component of *backward-looking* responsibility. Next, we focus on Mark Bovens (2007) notion of "mechanisms of accountability" and thereby identify three possible accountability gaps for Autonomous Weapon Systems based on the three layers identified by Van den Berg (2015). In Sect. 3, we show how accountability and control are linked and provide an engineering, socio-technical and governance perspective on control. We describe how combining these three perspectives constitute the notion of Comprehensive Human Oversight. Section 4 presents the Comprehensive Human Oversight Framework, consisting of three horizontal and three vertical layers (nine blocks) and their connections. In Sect. 5, we apply the Dutch military control instruments that are currently used to the nine blocks of the Comprehensive Human Oversight Framework. Section 6 shows the applicability of the Comprehensive Human Oversight Framework to Autonomous Weapon Systems. This reveals at least two gaps in the control mechanisms in relation to Autonomous Weapon Systems. In Sect. 7 we conclude by identifying three first options for the design of a control mechanism in the technological, the socio-technical and the governance layer as future work in order to achieve Comprehensive Human Oversight and ensure accountability over Autonomous Weapon Systems.

[1] With the exception of Santoni de Sio and Van den Hoven (2018).

2 From Responsibility to Accountability

In this section we first define the notion of accountability in relation to the broader concept of moral responsibility. Next, we describe accountability gaps or vacuums and conclude by presenting three possible accountability gaps for Autonomous Weapon Systems.

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 or 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. 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.

The definitions of Van de Poel (2011) above describe a form of responsibility that is *virtue-based responsibility*. A second form is of responsibility is *accountability*

in the way that X is accountable for Y for Z. In this sense accountability performs functions of scrutiny, for example calling someone to account, requiring justifications and imposing sanctions (Mulgan 2000).

These forms of responsibility are conceptually and casually related in many ways. For instance, once 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 (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 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* as *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.

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, under review). However, whereas the relationship between control and blameworthiness has been widely studied in philosophy (Fischer and 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.

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

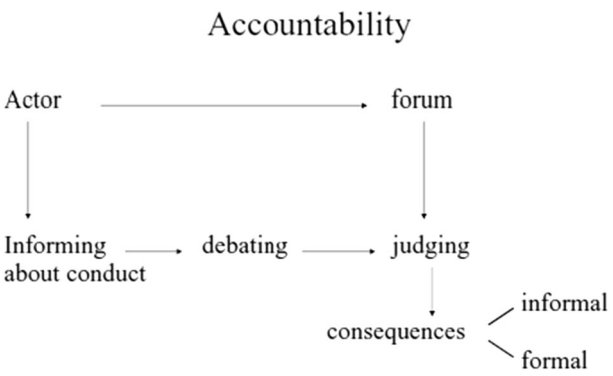

Fig. 2 Elements of accountability concept (as in: Bovens 2007)

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 et al. 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 et al. 2016). Figure 2 provides an overview of the relationship between the accountability elements. Accountability is not only scrutiny after the event has occurred, it has also 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 and 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 and 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 full accountability mechanism because the possibility of judgement and sanctions are lacking, and the relationship between the actor and forum is not clearly described.

Many scholars point to gaps in accountability relationships that will occur in the deployment of Autonomous Weapon Systems. 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 Systems 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-responsibility, instead of only backward-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 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 (2006) 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 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.

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 layers:

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 there aren't sufficient mechanisms for accountability in place.
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, sociotechnical and governance perspective based on the layers described by Van den Berg (2015) 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 are 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.

3 From Accountability via Control to 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.*' We have seen above that the concept of "Meaningful Human Control" was introduced in the debate on the ethics of Autonomous Weapon Systems, among other things to preserve human accountability. 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 toward Comprehensive Human Oversight mechanisms.

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).

3.1 Engineering Perspective on Control

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 plants

control (Åström and Kumar 2014; Pigeau and McCann 2002). In general, a control system has four common characteristics: (1) it is a *dynamic system* with responses that evolve in time and have 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.

3.2 Socio-technical Perspective on Control

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 has delegated 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 non-human intelligent systems, like Autonomous (Weapon) Systems.

3.3 Governance Perspective on Control

The governance perspective on control describes which institutions or forums supervise the behavior 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 coming from the philosophical debate on free will and moral responsibility. 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 persons who design or interact with the system (Santoni de Sio and 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. We may call this broad Meaningful Human Control. However, this wider conception of the control loop 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.

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 et al. 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.

Bovens (2007) notices 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, 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 section, we present the Comprehensive Human Oversight Framework for Autonomous Weapon Systems by describing its composition of layers and columns and their connections. We also show where gaps in the control mechanisms of the Comprehensive Human Oversight Framework exist.

4 A Comprehensive Human Oversight Framework for Autonomous Weapon Systems

Combining our analysis of accountability and perspectives on control, with the three layers described used by Van den Berg (2015) to characterize the cyber domain, in the phases of weapon deployment of an Autonomous Weapon System, results in a Comprehensive Human Oversight Framework depicted in Fig. 3. In this section we present the Comprehensive Human Oversight Framework by describing the layers and the connections between them and identifying gaps in the control mechanisms.

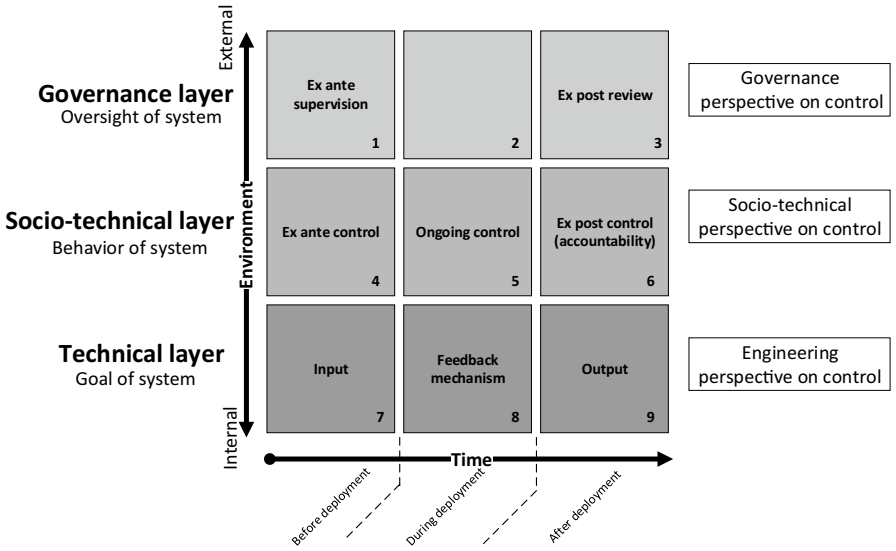

Fig. 3 Comprehensive Human Oversight Framework

The Comprehensive Human Oversight Framework consists of three horizontal layers that are based on the three-layered model that Van den Berg (2015) created to describe cyber space. These layers can also be linked to the accountability mechanisms and control perspectives described, respectively in Sects. 2 and 3. On the x-axis *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 concept 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.

The figure above 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.

4.1 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 from the human operator (block 7), the system's feedback mechanism should be robustly and verifiably 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 and the processes behind them (block 9).

4.2 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). During the use they 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). After deployment they should be able to inspect and assess the behaviour of the system to be able to account for its actions (block 6).

4.3 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 systems 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 is 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 the Fig. 3 a clear gap is visible in the governance layer of the middle column. A mechanism in block 2 is 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. In turn this affects 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.

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. We conclude by describing the connections and feedback loop between the layers. We recommend to close the feedback loop in the governance layer to incorporate the findings of the review process in the mandate for a next mission.

5 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 and 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 and 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 Sect. 3, we turned to the military domain to identify the military control instruments that are currently used in the three layers. We found that in the military domain there is a control mechanism in each layer before, during and after deployment of a weapon system. For each block examples of these mechanisms in the Netherlands are plotted in the Comprehensive Human Oversight Framework (Fig. 4) and described below.

The military control instruments in the Comprehensive Human Oversight Framework per block in the Netherlands 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

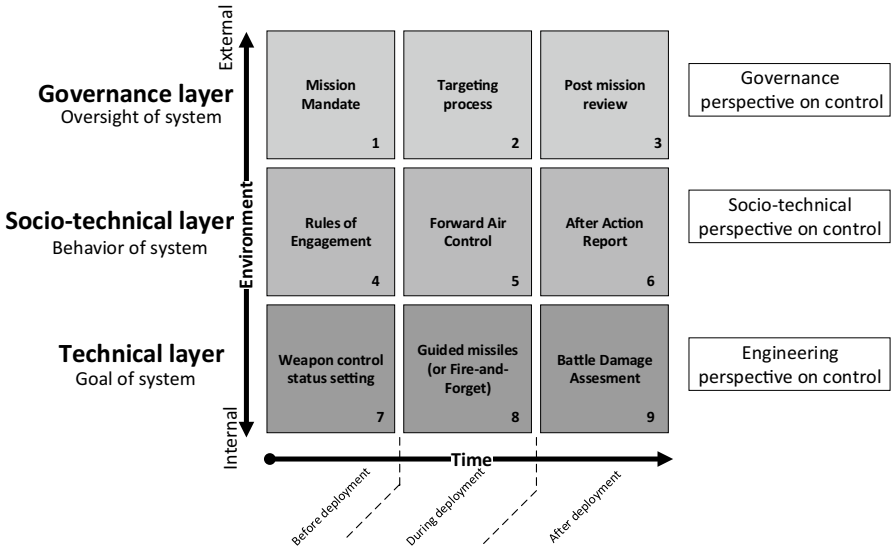

Fig. 4 Military control instruments

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 int 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.
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, determining the Area of Responsibility (AOR) and the targeting process.
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.

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 Fig. 3, the military domain has an oversight mechanism during deployment in block 2 (see Fig. 4). 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 Fig. 4 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 Sect. 4, which is based on the literature, to the application of the Framework to Autonomous Weapon Systems. 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 Sect. 2.

6 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* which is a not well-defined and often misunderstood concept. Castelfranchi and Falcone (2003) define autonomous 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 and Falcone 2003).

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 and CAVV 2015) 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 et al. 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.

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:

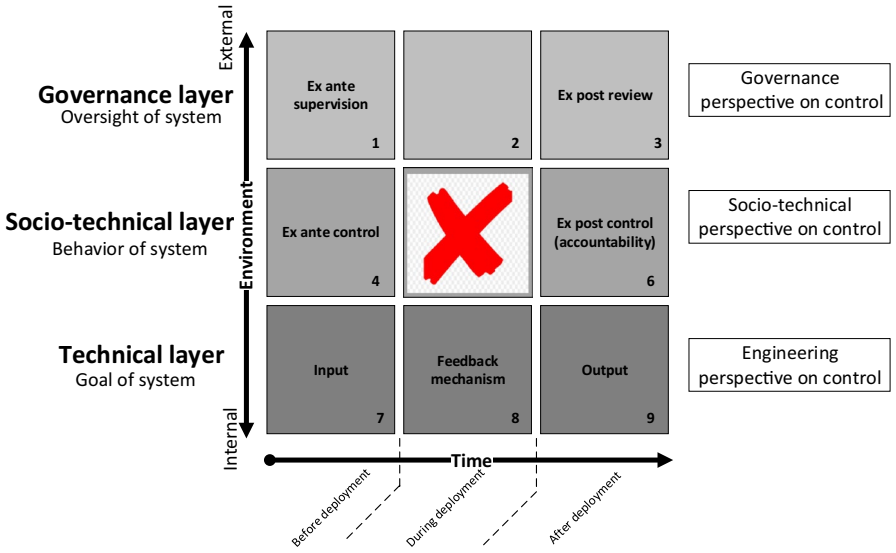

Fig. 5 Comprehensive Human Oversight Framework for Autonomous Weapon Systems

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 (y) (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 Sect. 2 on MHC above).

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 Fig. 5).

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 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 and Van den Hoven 2018), as control mechanisms in the governance layer.

Comparing the Comprehensive Human Oversight Framework in Fig. 3 to that of Autonomous Weapon Systems in Fig. 5 reveals two gaps in the control mechanisms that can be linked to the accountability gaps in Sect. 2: (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.

7 Conclusion

The notions of accountability and responsibility are key concepts in the societal and academic debate. Accountability can be regarded as *backward-looking* responsibility to account for conduct after actions have occurred. However, many point to the fact that Autonomous Weapon Systems may lead to an *accountability gap*, *accountability vacuum* or an erosion of accountability relationships. We have identified three possible accountability gaps on three different layers: (1) a technical accountability gap, (2) a socio-technical accountability gap and a (3) a governance accountability gap. Accountability is a form of control and the notion of control can be viewed from different perspectives. In this paper we mention 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 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. Future work will have to address these gaps to assess whether other gaps may emerge at other levels which need to be filled in order to ensure accountability over Autonomous Weapon Systems.

7.1 Future Work: Designing a Mechanism for Control to fill the Control Gaps

The Comprehensive Human Oversight Framework for Autonomous Weapon Systems highlights at least two main gaps that can be linked to the accountability gaps described in Sect. 2. This raises the issue if this is sufficient for control to be meaningful for the deployment of Autonomous Weapon Systems. It seems that this is not the case and this deficiency indicates a need for additional mechanisms for the deployment of Autonomous Weapon Systems. In future work we will study which mechanism in which layer will lead to sufficient human oversight over the deployment of Autonomous Weapon Systems. We have identified three first options for the development of such a mechanism (see Fig. 6): (1) a monitoring process in block 2 to ensure oversight of weapon system, (2) a mechanism in block 5 of the socio-technical layer, or (3) a mechanism in block 8 of technical layer to control the goal of the system. For the first option several directions could be taken for further research,

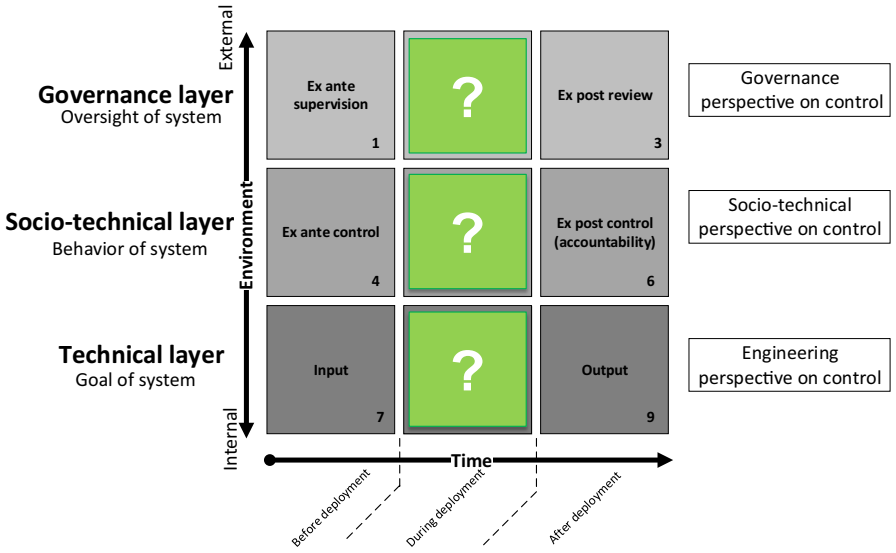

Fig. 6 Future Comprehensive Human Oversight Framework for Autonomous Weapon Systems

for example looking at mechanisms in the policy or organisational literature. For the second option the concept of broad Meaningful Human Control could be used to study if this fills the socio-technical gap in block 5. The third option is often mentioned in the policy and engineering domain as solution to fill the gap in the socio-technical layer. All three options will be studied in the next phase of our research. Given the many interconnections between various components of control it is to be expected that more changes in other blocks (ex ante and/or ex post) may be needed to achieve the desired results in the "during deployment" blocks. Future work will also explore these possible additional changes.

Another possible direction for future research is to evaluate the Comprehensive Human Oversight Framework in other fields where autonomous systems are used. For example, in the case of Autonomous Vehicles, 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.

References

AIV and 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

Alberts, D. S., Hayes, R. E. (2006). *Understanding command and control*.

Alston, P. (2010). *Study on Targeted Killings', Report of the UN Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, UN Doc A*.

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*., Robot law Cheltenham: Edward Elgar Publishing.

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

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. *European Law Journal, 13*(4), 447–468.

Bovens, M., Schillemans, T., & Goodin, R. E. (2014). Public accountability. *The Oxford Handbook of Public Accountability, 1,* 1–20.

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.

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

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., & Falcone, R. (2003). *From automaticity to autonomy: the frontier of artificial agents* (pp. 103–136)., Agent autonomy Berlin: Springer.

Crootof, R. (2015). War torts: Accountability for autonomous weapons. *University of Pennsylvania Law Review, 164,* 1347.

Cummings, M. L. (2006). Automation and accountability in decision support system interface design.

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)*.

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

Fischer, J. M., & Ravizza, M. (1998). Responsibility and control: A theory of moral responsibility.

Fitzsimmons, S., & Sangha, K. (2010). Killing in high definition. *Technology, 12,* 289–292.

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

Gardner, J. (2007). The mark of responsibility. *Offences and defences* (pp. 177–200). Oxford: Oxford University Press.

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

Greer, S. L., Wismar, M., Figueras, J., & McKee, C. (2016). Governance: a framework. *Strengthening Health System Governance, 22,* 27–56.

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

Hulstijn, J., & Burgemeestre, B. (2014). *Design for the values of accountability and transparency* (pp. 1–25)., Handbook of ethics, values, and technological design: sources, theory, values and application domains Berlin: Springer.

ICRC. (2018). *Ethics and autonomous weapon systems: An ethical basis for human control?* Retrieved 31 January 2020 from Geneva. https://www.icrc.org/en/download/file/69961/icrc_ethics_and_autonomous_weapon_systems_report_3_april_2018.pdf

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

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

Koppell, J. G. (2005). Pathologies of accountability: ICANN and the challenge of "multiple accountabilities disorder". *Public Administration Review, 65*(1), 94–108.
Liao, S.-H. (2008). Problem structuring methods in military command and control. *Expert Systems with Applications, 35*(3), 645–653.
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.
Mecacci, G., Santoni De Sio, F. (2019). Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics and Information Technology*.
Meloni, C. (2016). *State and individual responsibility for targeted killings by drones* (p. 47). Drones and Responsibility: Legal, philosophical and socio-technical perspectives on remotely controlled weapons.
Mulgan, R. (2000). 'Accountability': An ever-expanding concept? *Public Administration, 78*(3), 555–573.
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
Pelizzo, R., Stapenhurst, R., Olson, D., Parliamentary Oversight for Government Accountability. (2006). Research collection school of social sciences. Paper 137. https://ink.library.smu.edu.sg/soss_research/137
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*. Toronto: Defence and Civil Institute of Environmental Medicine.
Radin, B. A., & Romzek, B. S. (1996). Accountability expectations in an intergovernmental arena: The national rural development partnership. *The Journal of Federalism, 26*(2), 59–81.
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., 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 Au-Tonomous Weapons Systems, UN Convention on Certain Conventional Weapons.
Romzek, B. S., & Dubnick, M. J. (1987). Accountability in the public sector: Lessons from the challenger tragedy. *Public administration review, 1,* 227–238.
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.
Schedler, A. (1999). *Conceptualizing accountability* (p. 14)., The self-restraining state New York: Power and accountability in new democracies.
Scott, C. (2000). Accountability in the regulatory state. *Journal of Law and Society, 27*(1), 38–60.
Sparrow, R. (2016). Robots and respect: Assessing the case against autonomous weapon systems. *Ethics & International Affairs, 30*(1), 93–116.
UN GGE LAWS. (2018). *Emerging Commonalities, Conclusions and Recommendations*. https://www.unog.ch/80256EDD006B8954/(httpAssets)/EB4EC9367D3B63B1C12582FD0057A9A4/$file/GGE+LAWS+August_EC,+C+and+Rs_final.pdf
Van de Poel, I. (2011). *The relation between forward-looking and backward-looking responsibility* (pp. 37–52)., Moral responsibility Berlin: Springer.
Van den Berg, J. (2015). Wat maakt cyber security anders dan informatiebeveiliging? *Magazine Nationale Veiligheid en Crisisbeheersing, 2,* 2015.
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.
Williams, G. (2008). Responsibility as a virtue. *Ethical Theory and Moral Practice, 11*(4), 455–470.

Affiliations

Ilse Verdiesen[1] · **Filippo Santoni de Sio**[1] · **Virginia Dignum**[1,2]

✉ Ilse Verdiesen
e.p.verdiesen@tudelft.nl

[1] Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands

[2] Umeå University, 901 87 Umeå, Sweden