Philosophy & Technology (2023) 36:3
https://doi.org/10.1007/s13347-022-00602-7

RESEARCH ARTICLE

The Responsibility Gap and LAWS: a Critical Mapping of the Debate

Ann-Katrien Oimann[1,2]

Received: 24 May 2022 / Accepted: 21 December 2022 / Published online: 5 January 2023

Abstract

AI has numerous applications and in various fields, including the military domain. The increase in the degree of autonomy in some decision-making systems leads to discussions on the possible future use of lethal autonomous weapons systems (LAWS). A central issue in these discussions is the assignment of moral responsibility for some AI-based outcomes. Several authors claim that the high autonomous capability of such systems leads to a so-called "responsibility gap." In recent years, there has been a surge in philosophical literature around the concept of responsibility gaps and different solutions have been devised to close or bridge these gaps. In order to move forward in the research around LAWS and the problem of responsibility, it is important to increase our understanding of the different perspectives and discussions in this debate. This paper attempts to do so by disentangling the various arguments and providing a critical overview. After giving a brief outline of the state of the technology of LAWS, I will review the debates over responsibility gaps using three differentiators: those who believe in the existence of responsibility gaps versus those who do not, those who hold that responsibility gaps constitute a new moral problem versus those who argue they do not, and those who claim that solutions can be successful as opposed to those who believe that it is an unsolvable problem.

Keywords Responsibility gap · Artificial intelligence · Moral responsibility · AI ethics · Command responsibility · Autonomous weapons systems

✉ Ann-Katrien Oimann
Ann-katrien.oimann@mil.be; ann-katrien.oimann@kuleuven.be

[1] Department of Behavioral Sciences, Royal Military Academy, Brussels, Belgium

[2] Department of Philosophy, KU Leuven, 3000 Leuven, Belgium

1 Introduction

In developing modern weaponry, people are constantly looking for ways to generate maximum damage to the target while minimizing the risk for the operator (Ohlin, 2017). In line with this, there has been a rise in the use of semi-autonomous systems and research into fully autonomous systems (Egeland, 2016; Hellström, 2013). This has led to debates about the ethical use of lethal autonomous weapon systems (LAWS) in the highest circles at national and international level.[1] The international community has focused extensively on the question of whether LAWS will be able to comply with the rules of International Humanitarian Law (IHL). This is especially true of the *jus in bello* requirements of distinction, proportionality, and necessity. Critics of the use of LAWS fear that the systems will be indiscriminate with regard to combatants and non-combatants and that such systems are unable to adequately weigh the military advantage of an attack against the damage because these evaluations are to a large extent context-dependent and thus difficult to determine numerically (Asaro, 2012; Dremliuga, 2020; Egeland, 2016; Van Severen & Vander Maelen, 2021).

The solutions of these and other problems depend on future technological developments. Once the technology meets the required thresholds in humanitarian law, there is arguably no further legal obstacle to its future use. However, the possible future use of LAWS raises another ethical problem related to the autonomous character of the technology itself: the problem of assigning moral responsibility for AI-based outcomes. In recent years, both in the legal sphere and in philosophy, attention has been paid to the difficulty of allocating moral responsibility for errors made by LAWS. Some authors argue that the increasing level of autonomy in weapon systems will lead to a "responsibility gap" (de Jong, 2020; Matthias, 2004; Roff, 2014; Sparrow, 2007).[2] According to this view, it is impossible to identify anyone who can be held responsible for harm caused by LAWS. The reason for this is, on the one hand, that it would be unfair to hold humans responsible as they no longer control the system (due to its high degree of autonomy and capacity for self-learning), while, on the other hand, it is impossible to hold the system itself responsible as it has no consciousness and cannot be the addressee of punishment or other forms of blame.

[1] See among others: the open letters in 2015 and 2017 by renowned technology experts about the dangers of LAWS; The Campaign to Stop Killer Robots (a global coalition of 172 NGOs in 65 countries) calling for a new international treaty to ensure weapons are always controlled by humans; Resolution in 2018 by the Belgian Parliamentary Defense Committee to prohibit the use of LAWS by the Belgian army as well as the production by arms manufacturers in Belgium; Resolution by the European Parliament in 2021 calling for an EU strategy against LAWS and to prohibit so-called killer robots; meetings organized by the Group of Governmental Experts (GGE) on emerging technologies in the area of lethal autonomous weapons systems of the Convention on Certain Conventional Weapons (CCW) in the UN and various speeches by UN Secretary-General António Guterres (most recent one in May 2020 on the protection of civilians in armed conflicts).

[2] The term was first used by Andreas Matthias with respect to autonomous machines (2004) and was later applied to autonomous weapon systems by Robert Sparrow (2007)

In the case of bad outcomes caused by LAWS, a distinction can be made between so-called easy cases and hard cases. Examples of easy cases would be the following: a software engineer who has intentionally programmed a weapon to target civilians, or a human operator who deployed the weapon to carry out unlawful attacks. In easy cases, someone (it may be a programmer, a manufacturer, or a user (Pagallo, 2013, 69)) exploits a system as a tool to commit a certain crime. In these cases, they will be held responsible (Saxon, 2016). In hard cases, harm is caused by LAWS, yet no human acted intentionally or carelessly (Königs, 2022, 7; McDougall, 2019, 70; Simmler & Markwalder, 2019, 7–9; Crootof, 2016, 1377). In these cases, there is a responsibility gap if no human involved can or should be held responsible. In recent years, there has been a surge in philosophical literature around the concept of responsibility gaps and various positions have been taken.[3] The concept also crops up frequently in legal literature, often under the term "accountability gap."[4] Considering the vast differences in assumptions in these debates, it can be difficult to determine how the views relate to each other, in what ways they are compatible, and in what exact points they differ.

In order to move forward in the research around LAWS and the problem of responsibility, it is important to increase our understanding of the different perspectives and discussions. This paper attempts to do so by disentangling the various arguments and providing a critical overview. It is primarily intended as an ethical analysis, but the paper will also build on and discuss relevant legal literature, since in many of the discussions on whether the autonomous power of systems would make it impossible to hold anyone responsible, moral and legal responsibility are often taken together.[5] To fully understand the debate and to explain how the various interlocutors reach their disparate conclusions regarding the presence or absence of responsibility gaps, it is useful to first have a good understanding of the technology. This will be done by giving a short overview of the state of the technology of LAWS (Sect. 1). Next, I will examine the debates around responsibility gaps with respect to LAWS, with the aim of providing clarity as to the multitude of prevailing views. I will do this by using three differentiators: those who believe in the existence of responsibility gaps versus those who do not (Sect. 2), those who hold that responsibility gaps constitute a new moral problem versus those who argue they do not (Sect. 3), and those who claim that solutions can be successful as opposed to those who believe that it is an unsolvable problem (Sect. 4).

[3] The issue has been discussed by a number of authors, some of the most recent ones: (Champagne & Tonkens, 2015; Chengeta, 2016; Crootof, 2016; Danaher, 2016, 2022; de Jong, 2020; Nyholm, 2018; Santoni de Sio & Mecacci, 2021; Tigard, 2020).

[4] Accountability and responsibility do not completely overlap, as agents can be morally responsible without being accountable and vice versa, but accountability and liability often presuppose moral responsibility. This is especially the case in criminal law, because although some moral wrongs do not concern criminal law, criminal law generally does deal with moral wrongness as criminal law is a practice that holds people responsible for wrongs they have committed.

[5] See for example: (Danaher, 2022; Santoni de Sio & van den Hoven, 2018; Roff, 2014). Throughout the rest of this paper, I will therefore refer to moral responsibility and only when strictly necessary use accountability or liability. When referring to other authors, I will use their original terminology.

2 State of the Art LAWS

Throughout history, new weapons technologies have significantly impacted the way people conduct war. With the discoveries and improvements within the field of AI, and particularly the second generation of AI systems,[6] the possibility of LAWS came into view. Despite various attempts by the Convention on Certain Conventional Weapons (CCW), there is still no universal definition of such systems, so several definitions are currently in circulation, each with its own characteristics and emphases.[7] While there are many differences, most understand LAWS to mean the following: "systems that once activated can select and engage targets without further intervention by a human operator."[8] These kind of systems are distinguished from semi-autonomous systems where humans still select the targets.[9] In order to clarify the distinction, a division is often made between systems with humans in, on or out of the loop.[10] I will briefly discuss and use this subdivision in the following paragraph to provide an overview of some of the current technologies and their underlying differences.

Systems with a human "in the loop" are the conventional systems that are remotely controlled such as unpiloted aerial vehicles (UAV) or unpiloted ground

[6] Second generation systems can be roughly described as statistical learning models, a form of AI that incorporates machine learning.

[7] For an overview of the different definitional approaches and the discrepancies between proposed definitions by countries see the report by the UNIDIR: (*The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches*, 2017).

[8] ICRC's working definition; Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, Christof Heyns, UN doc A/HRC/23/47, para 38; US Department of Defense, 'Directive 3000.09 Autonomy in Weapon Systems', 21 November 2012, p. 13; B. Docherty, Losing Humanity: The Case against Killer Robots, Human Rights Watch, November 2012, p. 2; Yoram Dinstein, 'Autonomous Weapons and International Humanitarian Law', in *Dehumanization of Warfare*, ed. Wolff Heintschel von Heinegg, Robert Frau, and Tassilo Singer (Cham: Springer International Publishing, 2018), 15–20 (p. 17), https://doi.org/10.1007/978-3-319-67266-3_2

[9] A clear definition of semi-autonomous systems can be found in the U.S. DoD Directive 3000.09, where they are being described as follows: "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. This includes (a) semi-autonomous weapon systems that employ autonomy for engagement-related functions, including, but not limited to, acquiring, tracking, and identifying potential targets; cueing potential targets to human operators; prioritizing selected targets; timing of when to fire; or providing terminal guidance to home in on selected targets; provided that human control is retained over the decision to select individual targets and specific target groups for engagement (b) "fire and forget" or lock-on-after launch homing munitions that rely on TTPs to maximize the probability that only the targets within the seeker's acquisition basket when the seeker activates are those individual targets on specific target groups that have been selected by a human operator.".

[10] It should be noted that this division is contested because of disagreement over the scope of the loop. Some argue that the debate is too limited to selection and engagement and that the loop should be understood more broadly, as humans remain involved in the overreaching goals regarding the design and the deployment of the system and continue to play an important role regarding the rules of engagement. The USA stated in this regard that "there are no fully autonomous systems just as there are no fully autonomous sailors, airmen or marines", US DoD Science Board 2012 'Task force report: role of autonomy in DoD systems', p. 23–24. See in this regard also: Council of Europe study DGI(2019)05, "Responsibility and AI," p. 20.

vehicles (UGV). In these systems, data is collected and processed that serves as input for the decision-making process. However, it is the human operator who selects the targets and maintains direct control over the engagement process. Systems with a human "on the loop" include counter rocket, artillery, and mortar systems (C-RAM), such as the Iron Dome.[11] In these types of systems, humans only perform supervisory tasks, but they are able to intervene when necessary. The renowned SGR-A1 weapon[12] used by South Korea in the demilitarized zone or the Super aEgis II[13] is also often classified as on-the-loop systems. Further mention can be made of Israel's Harop.[14] This is a loitering munition that searches within a certain geographical area for targets that meet certain criteria and eliminates them if found. Finally, there are the autonomous systems where humans are completely "out of the loop." Until recently, there has been a broad consensus that systems which can select and engage human targets in a dynamic environment without human intervention do not yet exist.[15] However, a recent report by the UN Panel of Experts on Libya points to the use of a LAWS, the STM Kargu-2, which may have hunted down and attacked retreating soldiers last year in Libya without data connectivity between the operator and the system.[16] Kargu-2 is a loitering drone that classifies objects and makes decisions based on machine learning and real-time image processing.

Debates around LAWS often run high. Several parties advocate a full preventive ban on LAWS without exception,[17] while other countries like the USA, Russia, and Israel consider that the current IHL framework is sufficient. What makes autonomous weapons distinctive, and why the reported use of a system like Kargu-2 generates so much debate, can be explained by a combination of factors. A first factor is the use of an autonomous weapon *beyond the non-critical areas* such as transport, logistics, navigation, and surveillance. A high degree of autonomy is already

[11] These systems can detect and destroy incoming rockets, artillery, and mortar shells in the air before they hit their targets.

[12] The system is developed by Samsung Techwin and Korea University to replace South Korean human guards along the DMZ to detect North Korean soldiers. It is unclear how exactly the system is used in practice and whether the system first warns an operator before firing. In that case, the final decision would be made by humans. However, in theory, the system can eliminate targets without human intervention.

[13] The Super aEgis II is a remote controlled weapon station manufactured by the South Korean DoDaam that can autonomously detect, track, and target humans up to 3 km away: http://www.dodaam.com/eng/sub2/menu2_1_4.php Although the system can theoretically fire automatically, the company stated that all its customers require the entering of a password by a human operator before firing: https://www.bbc.com/future/article/20150715-killer-robots-the-soldiers-that-never-sleep

[14] The system is developed by Israel Aerospace Industries: https://www.iai.co.il/p/harop. The weapon was reportedly used in combat in the Nagorno-Karabakh conflict in 2016 by Azerbaijan.

[15] See for example: Marco Sassòli, "Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to Be Clarified," *International Law Studies* 90 (2014): 308–40.

[16] UN Security Council's Panel of Experts on Libya, "Final report of the Panel of Experts on Libya established pursuant to Security Council resolution 1973 (2011)," p. 1–48 (p. 17): https://documents-dds-ny.un.org/doc/UNDOC/GEN/N21/037/72/PDF/N2103772.pdf?OpenElement.

[17] Among others: Campaign against Killer Robots (a coalition of NGOs), a list of 30 countries in the UN, and numerous scientists. For an overview see: https://futureoflife.org/2018/06/05/lethal-autonomous-weapons-pledge/

common in reconnaissance systems, but in the case of LAWS, it involves delegating the decision to eliminate a target. A second factor is the potential for *offensive use*, since most automatic systems today are only used defensively. Current systems are mostly used to counter incoming danger and do not actively search for potential targets. Thirdly, there is the *lethal* aspect. Most automatic systems are used anti-materially and can cause collateral damage, but they are not specifically designed to eliminate human opponents.[18] The fourth factor has to do with the use of *machine learning* algorithms. These algorithms differ from rule-based systems that are pre-programmed in the form of the logical rules. By contrast, a machine learning system improves its performance on a specific task based on experience, i.e., past input. This last point poses a serious challenge in the context of LAWS. A large amount of accurate data is required for the system to be able to distinguish the right targets in different environments and circumstances, especially given that these are black-box systems with self-learning capacity (Boulanin & Verbruggem, 2017, 25). Another difficulty with regard to machine learning in a warfare context is defining what the task of the system exactly is and how optimization should be determined.

Taken separately, the factors are not considered to cause moral problems, and even most combinations of two or three factors do not trigger major concern in international discussions. For example, while the SGR-A1 is a lethal system capable of eliminating human targets by means of a thermographic camera and a laser rangefinder, it is only used in a well-defined specific zone and its task is to eliminate everything that enters that zone. The Iron Dome is capable of eliminating targets without human intervention but has a strictly defensive anti-material use. Compared to other existing weapon systems with some degree of autonomy, the combination of the four factors mentioned above raises significant concerns. Systems with a high degree of autonomy have so far been used mainly in demarcated areas or in areas with less chance of obstacles, such as at sea and in the air. An urban environment such as a city or village with many people and therefore a high probability of changes seems much less suitable for the deployment of such systems. Furthermore, the requirement for a high degree of precision is also an obstacle for the system. A system that needs to recognize a certain object that has a high degree of uniformity and always comes from the same direction is easier to develop than a system that is tasked to distinguish between people. This is further complicated in the case of individual targeting or situational targeting where the identification of enemies cannot be done solely based on certain distinctive signs but must be inferred from the role of a particular individual vis-à-vis the hostilities or from the alleged behavior (Margulies, 2019, 409). The most advanced systems we have known so far were only capable of performing relatively simple tasks in relatively simple environments. In the longer term, LAWS seems to be able to overcome this paradigm.

[18] Exceptions are the SGR-A1 and the Super aEgis II in the DMZ, but these systems only operate in a selected area and human intervention is still possible.

3 The Existence of Responsibility Gaps

In order to move forward in the research and debates around the problem of responsibility gaps for LAWS, it is important to increase our understanding of the different perspectives and positions in the literature. This paper attempts to do so by disentangling the various arguments and providing a critical overview. The first significant distinction is between the authors who believe that responsibility gaps exist and those who believe that they do not. I argue that this distinction arises due to three factors: disagreement about the object of application of the concept of responsibility, the confusion with the problem of many hands, and the disagreement about the nature of LAWS. I will discuss these one by one.

The first disagreement about the existence of responsibility gaps stems from disagreement about the object of application of the concept of responsibility.[19] In cases where something goes wrong, different scopes of responsibility can be distinguished from each other. These include responsibility for the development, design, production, proliferation, deployment, and use of LAWS. Usually, the ways in which systems fail fit neatly in the abovementioned paradigm and moral responsibility can be attributed to the different actors. Traditionally, responsibility for the consequences of operations of machines is usually attributed either to the developers or to the users. This is solved by most legal systems as follows: if the system does not perform within the developers' specified parameters, then the fault is attributed to the developers. On the other hand, if the system works within the developers' specifications but the system is deployed in an unlawful manner, the users (understood in a broad sense)[20] will be responsible. Applied to a military context, these will be military commanders or political authorities. However, according to the authors who believe in the existence of a responsibility gap, this paradigm does not seem applicable in the case of LAWS and other intelligent systems (de Jong, 2020; Matthias, 2004; Roff, 2014; Sparrow, 2007). The reason is the increasingly autonomous nature of operating machines. The rules according to which the machines act are no longer all preprogrammed, and so the system can adapt itself. In other words, the potential consequences of the machines "actions" are no longer fixed during the production phase, but is changed during the operation of the machine itself, since the decisions are also based on data that the system has obtained from its surroundings and experiences (Galliott, 2020, 168; Matthias, 2004, 177). The use of AI and data-driven machine learning in decision-making decreases the possibility to ascribe

[19] As an anonymous reviewer has pointed out, this should be clearly distinguished from debates about disagreement on the concept of responsibility. Currently, philosophers disagree on what constitutes moral responsibility, and on necessary and sufficient conditions. This results in varieties of retrospective moral responsibility, including distinctions between attributability, accountability, and answerability. For an overview on the differences see: (Zimmerman, 2015). However, these debates should not detain us further here as they are not often addressed in the debates on responsibility gaps in LAWS. Exceptions in this regard are: (Santoni de Sio & Mecacci, 2021; Tigard, 2020).

[20] The term users should be understood in a broad sense because the process that leads to the use of LAWS involves decisions from different actors including commanders, operators, and political authorities.

responsibility to the human agents we normally hold responsible because they are unable to predict or control the outcome of the system's action. In summary, it is argued that developers (i.e., engineers, designers, and programmers) can no longer be held responsible because the system makes choices that could not have been predicted by the developers (Tollon, 2022; Egeland, 2016, 112; Sparrow, 2007, 70). Similarly, users such as commanders do not seem to be held rightfully responsible if the system is able to set its own targets (Sparrow, 2007, 71). At the same time, it is also impossible to attribute responsibility to the machine itself because it is generally assumed that they lack moral agency. In precise terms, proponents of the existence of responsibility gaps use the term in a narrow sense as they refer to the impossibility of attributing individual, moral, outcome (retrospective) responsibility to users for events caused by LAWS.

This is overlooked by some authors who have argued that there is no such gap in responsibility since it is possible in all cases to hold someone responsible. The authors that can be mentioned in this regard are as follows: Sebastian Köhler, Neil Roughley, and Hanno Sauer who argue that those who risked harm or made some minimal causal contribution can always be held responsible (2017) and Dante Marino and Guglielmo Tamburrini who mention that responsibility can be put upon computer scientists, engineers, or organizations based on prospective responsibilities (2020). What they overlook is the fact that those who believe in the existence of responsibility gaps use the term in a narrow sense and refer to a problem that exclusively refers to *ex post* responsibility for the users (broadly conceived) of such systems. The aforementioned different forms of responsibility (design, proliferation, use, etc.) coexist but are not complementary.[21] For example, if we look at traditional weapons technologies, we see that while the manufacturers are responsible for the safety of the weapon, this does not negate the responsibility of the operator for its use. The problem with the abovementioned authors is that they insufficiently acknowledging the blurring of the distinction between developers and users. In the case of LAWS, the distinction between developers and users blurs because some of the critical decisions about targeting that are made at the development stage, whereas in traditional weaponry, they are made exclusively by the users. This could possibly increase as in future scenario's military commanders could adapt the parameters of LAWS during deployment (Bo et al., 2022, 38). What makes it confusing is thus the fact that developers are not only to be considered for the attribution of responsibility for the design where the system works outside the pre-set parameters (as with traditional systems) but also occur among the range of subjects to be considered for the attribution of responsibility for the use of LAWS. The above authors claim, often on the basis of forward-looking responsibilities of certain actors, that there is no responsibility gap. However, when they refer to the various actors who could be held responsible, they seem to refer only to their responsibilities on design and

[21] In this sense, Thompson Chengeta has emphasized that "accountability forms of responsibility are not alternatives to the exclusion of the other (…) if an accountability gap is created in one form or mode of responsibility, it cannot be ignored on the basis that there are other persons who can be held responsible" (2016, 14).

lack an explanation on how they, as users, can be held responsible for bad outcomes involving LAWS. For it is true that we can always blame someone in the chain of command, as for example a software engineer, it should be proven how that agent can close the responsibility gap in the (narrow) sense used by the authors who worry about responsibility gaps as it remains unclear whether and how developers may be held responsible for bad outcomes that involve the LAWS they helped to develop.

A second disagreement about the existence of gaps in responsibility stems from the confusion with the problem of many hands. The problem of many hands is a term used to describe situations where many actors have contributed to an action that has caused harm and it is unclear how responsibility should be allocated.[22] A typical example is the case where an organization (government, private company, etc.) is responsible for an undesirable outcome, but where it appears that no member of the organization can be held responsible for this outcome. It is often used with respect to new technologies, because a large number of actors are involved in their development and use, and thus, there are many hands in the chain of responsibility.[23] It is important to keep in mind that the problem of many hands implies that it is very difficult or impossible to identify the right morally responsible agent, but it does not claim that there are no agents we could hold responsible. Some authors mention that responsibility gaps are caused or increased by the large amount of people involved in the life-cycle (Taylor, 2021, 324; de Jong, 2020). However, the problem of responsibility gaps is not related to the amount of people involved. There is no fixed amount of responsibility available for every outcome to be distributed among all those responsible for it, individual responsibility does not decrease as more people become involved.[24] The confusion stems from the fact that the problem of many hands is not only a practical-epistemic problem but also a normative one. The problem of many hands is often portrayed as a purely practical problem that can be solved by looking closely at the distribution of competence within the group and on that basis attributing the appropriate amount of moral responsibility. In essence, however, the problem of many hands is a normative problem so that even if someone had perfect knowledge of who causally contributed to what exactly, the problem could still not be solved (van de Poel et al., 2012, 61). This seems to imply that the problem of many hands leads to a situation where no one can be held responsible. However, this is not the case. The problem of many hands occurs in situations where our sense of justice holds the group responsible, but where this responsibility cannot be reduced to the responsibilities of the members of the group (de Lima & Royakkers, 2015, 117).[25] In these cases, the group is responsible without it seeming fair to

[22] The expression “many hands” was reportedly first used by Dennis Thompson in connection to officials in public administrations and later applied to computer technology by Helen Nissenbaum.

[23] The problem also occurs in non-technological areas such as public administration and climate change. For a comprehensive analysis, see: (Poel, 2015).

[24] For an analysis on this, see: (Kaiserman, 2021; Zimmerman, 1985).

[25] It is debated whether collective entities can be qualified as group agents that can be held morally responsible. Some authors consider groups to be “nonagential” (Taylor, 2021) or “minimal agents” (Himmelreich, 2019), while others describe them as “imperfect moral agents” (Crawford, 2013) and argue that they can qualify as responsible moral agents (Crawford, 2007; List, 2021). For the remainder of this article, it is not necessary to elaborate on this debate.

hold the members of the group responsible.[26] The crucial difference with the problem of the responsibility gap is that in traditional cases of many hands, it is still possible to designate a responsible agent, namely, the group.

A third disagreement about the existence of responsibility gaps derives from disagreement about the nature of LAWS. The question arises whether we should analogize LAWS more with conventional weapons or rather with human soldiers. The first analogy is used mainly by those who believe that there is no responsibility gap, while the second analogy is used by the authors who believe that there is a responsibility gap. According to the first view, LAWS should be considered tools and their decisions are merely delayed human decisions (Johnson & Axinn, 2013, 132). In this context, Marco Sassòli and Patrick Nagler argue that questions of responsibility in the case of LAWS should be treated in the same way as conventional weapons causing civilian casualties (Sassòli & Nagler, 2019, 527). Sassòli endorses the strict distinction between weapon systems and combatants: "The difference between a weapon system and a human being is not quantitative but qualitative; the two are not situated on a sliding scale, but on different levels—subjects and objects" (Sassóli, 2014, 323).[27] The rationale behind the analogy between LAWS and conventional weapons is that LAWS are essentially human-made automatic systems and not autonomous systems. Joanna Bryson accordingly states in this regard that autonomous systems are essentially non-existent and should be viewed as nothing more than tools (Bryson, 2010). Thus, according to these authors, there is no gap in responsibility as humans bear full responsibility for such systems. On the other hand, it is also often argued that advanced AI systems can no longer be seen as mere tools (Calo, 2015; Gunkel, 2020a; Lagioia & Sartor, 2020, 433). According to these views, the growing autonomous capability of certain systems means that the technology should not be seen as replacing the tools for the users, but as replacing the users themselves (Gunkel, 2020b, 310). Autonomous systems are similar to soldiers in the sense that they can take a certain action to achieve a predetermined state without any predefined rules. This means that they are no longer completely pre-programmed systems in which all steps are fixed in advance and the reasoning can be completely traced *ex post*, but systems with some discretionary power. The authors who defend the latter view therefore often point out that a responsibility gap arises because human beings can no longer be held fully responsible. In sum, disagreements about whether or not gaps in responsibility exist depend largely on how the author assesses the nature of LAWS and which analogy he or she uses.

[26] Sebastian Köhler, Neil Roughley, and Hanno Sauer have argued in this regard that this intuition is not true: "The fact that many people contributed to something that is morally significant isn't as such a problem for the ordinary conception. Second, even if each contribution is significantly small, this just implies that it is fitting to hold to account very many individuals in such cases. This, however, doesn't create a responsibility gap, but rather makes it appropriate to hold a large number of individuals to account to a relatively small degree each" (Köhler et al., 2017, p. 10).

[27] See in this regard as well Michael Robillard who argues that an autonomous weapon is either a socially constructed institution or it is a genuine agent: (2018).

4 The Responsibility Gap as a New Moral Problem

The second distinction that one can make is between theories that hold that the responsibility gaps in autonomous systems pose a new moral problem versus those who defend the view that they do not. Within the category of the authors who argue that responsibility gaps are not a new moral problem, we can roughly distinguish two positions: emphasizing the fact that gaps are not new and also occur in contemporary practice or arguing that gaps occur but should be seen as accidents, not as moral problems.[28]

The first position can be traced back to having a realistic view of the current practice of human decision-making in warfare. This view is clearly defended by Patrick Taylor Smith. His argument goes as follows: it is true that LAWS can cause unaccountable casualties, but these outcomes also routinely occur anyway. The pessimists about the solvability of responsibility gaps incorrectly assume that this outcome is unique when using LAWS. The use of LAWS may indeed pose a risk of LAWS acting in ways that commanders did not order or could not have anticipated, but this is not specific to LAWS as responsibility gaps also exist with human warfighters (Smith, 2019, 291). Dan Saxon makes a similar argument. He also points to the fact that such gaps in responsibility are not new, as they also occur in modern warfare: "ironically, commentators raise concerns about accountability gaps for autonomous drones when we tolerate similar gaps for other kinds of complex weaponry" (Saxon, 2016, 28).

The second strategy is more complex to understand because it finally explains away the problem of the responsibility gap. Here, it is argued that responsibility gaps should be classified purely as accidents. Sebastian Köhler argues that responsibility in human-AI interactions should be sought in the responsibility for the use of an instrument and treats it analogously to cases where we use and train non-human animals as instruments such as police dogs and racehorses (Köhler, 2020, 3134). On the one hand, these cases make it clear that it is impossible to completely eliminate harmful outcomes and that the person who failed to take the necessary precautions or who uses an instrument for a purpose that involves a risk of harmful consequences often remains responsible. On the other hand, according to Köhler, these cases also make it clear that there might occur situations where it is correct to think that no one is responsible since all duties of care have been taken. In these situations, it is inappropriate to speak of responsibility gaps, and one should rather consider these as accidents since they do not pose a moral problem. We find a similar line of reasoning in Thomas Simpson and Vincent Müller, but focused on LAWS. They argue that harmful effects due to LAWS should be compared to and treated like accidents with non-learning systems. What is decisive in both kind of cases is

[28] In this context, it is worth mentioning a remarkable recent proposal by John Danaher that goes a step further. He argues not only that responsibility gaps are not always problematic but also that there are sometimes reasons to welcome them. His position can be summarized as follows: there are times when we should prefer delegation to autonomous systems without attempting to resolve the responsibility gap that has arisen. See: (Danaher, 2022).

the so-called tolerance level. The tolerance level represents the minimum level of reliability that a system must achieve. They give the example of a bridge where the engineers must design a bridge that is sufficiently robust, and the contractors are then responsible for meeting that standard. In addition, there are parameters for the use of the bridge that users must adhere to (Simpson & Müller, 2016, 307). For all accidents that happen due to conditions within the required tolerance level, such as engineers who did not take into account strong temperature fluctuations or users who exceeded the maximum weight of the bridge with their vehicle, at least one person is responsible (be it the engineer, the controller, the user,…). But for all deaths that fall outside the required tolerance levels, however tragic, it is possible that no one is responsible (Simpson & Müller, 2016, 308). Take the example of a sudden rainstorm that normally occurs only once every 100 years and where it was determined that the bridge should not be able to withstand it because the probability was so low and the construction cost very high. Therefore, applied to the case of LAWS, if all necessary precautions have been taken, but an undesirable result still occurs, it should be considered an accident for which no one is responsible.[29] Consequently, to say that responsibility gaps are not problematic, because it is correct to say that no one is responsible, leads to explaining away the entire responsibility gap.

5 The Solvability of Responsibility Gaps

The third distinction is between the authors who claim that the gaps in responsibility can be closed as opposed to those who believe that this is impossible. Under the latter category can be placed both the fatalist[30] authors such as Robert Sparrow, Andreas Matthias, Heather M. Roff, and Roos De Jong and those who believe that these are merely (military) accidents (see *supra*). The strategies for resolving responsibility gaps vary widely. Apart from the solution in the previous paragraph, which holds that gaps in responsibility are purely (military) accidents and consequently solves the gaps by explaining them away, there are other genuine solutions possible. Broadly speaking, four can be distinguished: technical solutions, practical arrangements, holding the system itself responsible, and assigning collective responsibility. I will discuss these briefly.

5.1 Technical Solutions

The first strategy is to present the responsibility gap as a purely empirical problem that can be solved by tracing the causal chain through technical solutions. According to the authors who propose this solution, the main problem with responsibility gaps is the lack of transparency and explainability. As a result, once the so-called blackbox can be opened and we can identify every link between cause and effect, the

[29] For possible problems with this position and counter-arguments, see: (Santoni de Sio & Mecacci, 2021, 14–15).

[30] I borrow this term from (Santoni de Sio & Mecacci, 2021).

problem is solved.[31] Saxon goes even one step further, stating that the use of autonomous drones and the accompanying recording system may even eventually make it easier to establish individual criminal responsibility (Saxon, 2016, 34).

The problem is that these technical solutions are unable to address the real problem of the responsibility gap. In other words, it cannot fully grasp that the problem of responsibility gaps is a normative problem that concerns the (in)ability to assign individual moral outcome responsibility. The authors who solve responsibility gaps with technical solutions misunderstand the problem of the responsibility gap because they confuse the problem of attributing moral responsibility with a problem of causality. Admittedly, in some autonomous systems, there is a black-box problem, and it is difficult to trace bad outcomes back in time, since the performances of LAWS are the result of multiple decisions at multiple times. However, the technical solutions provided to gain more transparency in the cause-and-effect relationship can at best only identify the relevant causal agent(s). One could indeed argue that in the case of harmful effects of LAWS, there is less direct causal connection between the action of a human agent and the outcome since the moment we delegate a task (partially) to the system, there is a reduced causality between the giving of the order and the execution of the task. This is because some of the decision-making power has been transferred to a non-human agent who is no longer completely pre-programmed but has some discretionary power. In this case, while the ordering party still determines the top-level objectives, such as where and when the system will be deployed, the system can take certain actions to achieve a predetermined state, without any predefined rules. Yet, the blurring of the causal connection between an action and the outcome is not a substantial problem for assigning moral responsibility. Compare it to situations in ordinary hierarchical structures where subordinates have some degree of decision-making power. In such situations, although there is a reduced causality between the issuing of the command and the result, moral responsibility still flows upward in the chain of command. This demonstrates that the mere reduction of causal connection does not necessarily also reduce the attribution of moral responsibility. Consequently, tracing all the decisions that were made prior the occurrence of the conduct of LAWS is insufficient as a thorough solution since the problem of responsibility gaps cannot be found purely on the causal dimension.

5.2 Practical Arrangements

A second category of solutions includes those authors who want to solve the responsibility gap by making practical arrangements. Under this solution can be placed proposals to change liability regimes, such as the adoption of strict liability in criminal law, proposals that support the use of tort law or state responsibility in those cases,[32] the acceptance of so-called blank check liability where human agents, after

[31] However, Santoni de Sio and Mecacci have noted with respect to this solution that algorithmic explainability is neither a sufficient nor a necessary condition to address responsibility gaps, see: (Santoni de Sio & Mecacci, 2021, p. 16).

[32] See in this regard: (Amoroso & Giordano, 2019; Crootof, 2016; Santoro et al., 2008).

informed consent, hold themselves responsible for actions of military robots (Champagne & Tonkens, 2015), and accepting *ex ante* responsibility where human agents willingly take the “moral gambit” (Taddeo & Blanchard, 2022). It is not necessary for the purpose of this article to go into the specific nuances of each of these (largely legal) solutions, but it is sufficient to point out the underlying common denominator. In essence, they are all ways of correcting undesirable outcomes, regardless of whether there is moral culpability. In other words, these solutions are all a form of (forced) *taking* of responsibility. They are aimed at repairing harm and indemnifying the community against the costs of activities that could prove dangerous and pose a risk of serious harm. A concrete example of this would be the obligation for companies or governments involved in the development and production of LAWS to compensate victims for any resulting damages. The underlying idea is that the discussion of gaps in responsibility remains stuck in the language of moral culpability, but in situations where no individual acts intentionally, this is not a good solution and it is better to look at fault-without-guilt schemes to close gaps in responsibility.

This solution, however, is a purely practical response that, at best, leads to agreements on who should pay for the costs of the suffered harm but which cannot satisfy the victims’ feelings of resentment. Purely legal liability does not necessarily coincide with our human tendency for retribution. To fully understand this, we need to consider John Danaher’s concept of “retribution gaps.” Moral outcome responsibility is closely tied to retribution (van de Poel et al., 2012, 64). Danaher starts from empirical evidence suggesting that humans are innate retributivists: people tend to find someone to punish when morally harmful outcomes occur. Based on this, Danaher argues that increased robotization can lead to retributive gaps because there is a mismatch between certain psychological desires to punish and the lack of a suitable candidate (Danaher, 2016, 302). The proposals for practical arrangements to resolve gaps in responsibility cannot remedy this. Of course, we could “agree” to apply strict liability rules in cases where LAWS cause harm, but these civil legal standards cannot be used to address gaps in responsibility because the problem is not a lack of compensation but an inability to punish the right agent (Amoroso & Giordano, 2019; Chengeta, 2016). The problem of responsibility gaps must therefore be distinguished from “remedial gaps,” where it is only a matter of correcting bad situations (Taylor, 2021, 322). A thorough solution to responsibility gaps, on the other hand, involves something more: the ability to rectify situations through retribution.

5.3 *Holding the System Itself Responsible*

The following two solutions attempt to address this more fundamental problem. The third solution involves the possibility of holding the system itself responsible (List, 2021; Simmler & Markwalder, 2019; Tigard, 2021). The rationale is that, despite the fact that AI systems are developed by humans, the responsibility of AI-systems that have achieved a certain degree of autonomy cannot be reduced to human responsibility.[33] In this regard, Lagioia and Sartor argue that the assumptions used so far to exclude non-human entities from the scope of criminal law may need to be revised for AI systems. According to them, it appears that AI systems may not only satisfy the objective component, namely executing of the crime, but that the subjective component, the mental element, can also be attributed to certain AI systems under certain conditions (Lagioia & Sartor, 2020, 437).[34] We find a similar line of thought with Thomas Hellström. According to this author, autonomous power is the decisive factor in assigning moral responsibility to agents. In other words, the more power someone has, the more responsibility (s)he bears. Recent psychological research suggests that people assign moral responsibility to the robot and that the degree to which this happens is based on the degree of autonomy of the system (Furlough et al., 2021). As we will entrust more and more complex decisions to robots in the future, it seems that we will assign moral responsibility, shared with or separated from other agents, to the systems themselves (Hellström, 2013, 105). Daniel Tigard adds that in a sense it is possible to punish the system: "We can impose sanctions on artificial moral agent's domain of application, restrict its previously authorized behaviors, or work to rewrite any deviant or undesirable lines of code (…) While artificial moral agents cannot suffer like us, they can and should suffer the consequences of carrying out harmful behaviors. AI systems capable of functional morality might one day learn from and improve upon their unique mistakes, as a sort of reinforcement learning" (Tigard, 2021, 442–443).

This solution, I believe, is problematic because it does not accurately reflect the current nature of technology. I agree that LAWS are more than merely tools, but I reject the suggestion to treat them as genuine moral agents. Admittedly, the gap between LAWS and soldiers may be smaller than we initially tend to think. If we look at the different steps in the military decision-making cycle of a conventional air operation, it can be argued that the role of an operator in conventional air operations is also limited. Merel Ekelhof has shown, in her research on the current state of human control in military practice through an analysis of the military decision cycle in the case of a manned F-16 attacking a military base with GPS-guided weapons, that the primary role of the pilot is to navigate to the area in which the weapon can reach the target (Ekelhof, 2019). This is because the detailed mission planning is done by other air force personnel and the target is already validated prior to takeoff. The operator only needs to enter the target coordinates and position of the aircraft into the bomb's computer and press the switch to discharge the weapon since with GPS-guided munitions, it is not necessary to find the target visually and the computer suggests the most effective time to unload weapons. Then, the GPS-guided weapon navigates to the designated target coordinates to engage the target. As such, the operator has no active participation in either the planning phase or the targeting phase. Ekelhof further points out that under normal circumstances, it is by no means the case that an F-16 operator decides autonomously to attack a target (Ekelhof, 2019, 347). It seems that the increasing autonomous capability of LAWS would blur

[33] See in this regard (List, 2021, 1225).

[34] For an opposing evaluation, see: (Seher, 2016).

the fundamental distinction between weapons and combatants. This follows from the fact that a change in usage can be noticed, a growing number of systems are no longer used as tools but are for instance deployed to replace human border guards.[35] Furthermore, with the proliferation of various assistive devices, the role of the operator in conventional air operations has become increasingly limited. However, it is important to recall that LAWS are not created *ex nihilo*. Autonomous systems are capable of achieving some general goal without the possible solutions being narrowly defined since the system is able to learn new information, but its decision-making ability and autonomous power remain limited by the original programming of the software and by the hardware components.

5.4 Collective Responsibility

Finally, there is the solution of assigning responsibility to the humans involved based on collaborative nature of the agency. We find this view clearly held by Sven Nyholm. According to him, the gap in responsibility can be avoided by thinking in terms of human–robot collaborations rather than adhering to the idea that LAWS have some form of independent agency: "We should not think of the military robot as acting in an independent way. Rather, insofar as we attribute agency to it, we should think of it as exercising supervised and deferential collaborative agency. That is, we should think of it as collaborating with the humans involved and as being under the supervision and authority of those humans" (Nyholm, 2018, 1212). He illustrates the idea of collaborative agency with the example of a child gardening at the initiative of the parent, with the parent monitoring the child to make sure the child is doing the gardening in the right way. Just as the child does not act on his or her own initiative, neither do military robots act on their own initiative, since the actions of military robots are carried out based on human-initiated actions. Moreover, humans still exert some form of indirect control and oversight over the system, after all, if the system were to operate in an undesirable manner, the software would be modified, or its use discontinued. Nyholm therefore points out that in these cases, "there should be no question as to whether the humans involved in these collaborations bear a significant responsibility. Again, unless the robot appears out of thin air and starts acting in a wholly independent way within the human–robot interactions in question, it is collaborating with the humans involved" (Nyholm, 2018, 1213–1214). In summary, although it concerns a group-level action and the robot may be doing most of the work, human agents can and should be held responsible based on their role in the hierarchy, as they initiate and supervise the human–machine collaboration. Jai Galliott similarly argues that "All the involved agents and any others associated with the use of autonomous systems retain a share of responsibility, even though they may claim that they were not in complete or absolute control" (Galliott, 2020, 170). Both Nyholm and Galliott point out that the focus on individual agency is insufficient and they argue that we should think in terms of human–robot

[35] See *supra.*

collaboration. However, unlike Nyholm, Galliott's proposed theory of responsibility also explicitly includes the possibility of distributing part of the responsibility over non-human agents.

A number of the authors believe that the problem of assigning responsibility can be (partly) solved by looking at the hierarchical structure in the military (Himmelreich, 2019; Nyholm, 2018; Schmitt, 2012; Schulzke, 2013). As such, Nyholm argues that "When we try to allocate responsibility for any harms or deaths caused by these technologies, we should not focus on theories of individual agency and responsibility for individual agency. We should rather draw on philosophical analyses of collaborative agency and responsibility for such agency. In particular, we should *draw on hierarchical models of collaborative agency, where some agents within the collaborations are under other agents' supervision and authority*" (Nyholm, 2018, 1203). This refers to the method of accountability in a traditional military organization where responsibility flows upward within the chain of command (UNIDIR, 2017, 14). From the bottom up, soldiers are responsible to their commander for following (or not following) strict orders. Subsequently, it is the commander who is responsible for making decisions. In the event that something goes wrong, the commander cannot absolve him or herself of individual responsibility simply by referring to one's delegation to subordinates and so military leaders may be held responsible for crimes committed by their subordinates. In the legal literature, this solution, which is a concretization of collaborative agency, is better known under the doctrine of command responsibility. Command responsibility is a form of responsibility attribution whereby superiors can be held indirectly responsible for crimes committed by subordinates.[36] It is sometimes discussed as a solution to resolve gaps in responsibility because it allows moral agents to be held responsible for decisions they make about LAWS, while avoiding holding agents responsible who lack the ability to prevent the bad outcomes of LAWS.[37] While commanders may not have direct control over the actions of subordinates, they do have indirect control, including for the decision to relinquish part of the control over subsequent events to autonomous systems. Furthermore, commanders would still determine the general parameters under which the systems operate, such as where, when, how, and against whom military force may be used. As commanders still initiate and supervise the operation, it therefore seems plausible that even in the case of LAWS, the commanders remain responsible.

[36] The majority view in case law has been that the doctrine does not constitute a separate criminal offence but is a form of liability for omissions in relation to crimes committed by subordinates. See: ICTY Prosecutor v. Rasim Delić Case No. IT-04–83-T, para 59; (Mettraux, 2009, p. 18). However, there are also authors and judges who argue that the superior is not responsible for the same crimes as subordinates but for a separate crime of omission. For a clear explanation of the difference between the two views, see: (Sander, 2010).

[37] The authors that discuss command responsibility in relation to responsibility gaps in LAWS: (Bo et al., 2022, 35–38; Schwarz, 2021; McFarland, 2020, 162–164; Laura A Dickinson, 2019, 79–81; Margulies, 2019, 413–415; Nyholm, 2018; Saxon, 2016; Crootof, 2016, 1378–1381; Roff, 2014, 357–358; Schulzke, 2013).

This solution, however, runs the risk of insufficiently acknowledging the influence of the system's self-learning capacities on the hierarchical structure and, at worst, may result in the commander being unfairly held responsible solely on the basis of a particular position in the chain of command. Autonomous weapons differ from manually operated weapon systems since in the latter, humans select the objects of the attack and engage. With respect to LAWS, this is complicated because it will be the target selection code which runs in a LAWS control system that identifies and ultimately attacks the target. Autonomous systems are programmed to achieve some general goal, but the possible solutions are not narrowly defined and are affected by all possible interactions between the components of the system, chaotic and complex operating environments, and unpredictable actions of adversary parties (McFarland, 2020, 60). Therefore, it is practically unfeasible to predict the behavior of the weapon.[38] This is problematic because the autonomous power of LAWS could lead to the eroding of the commander's responsibility (Taddeo & Blanchard, 2022, 13–14). In order to hold commanders fairly responsible, it is necessary that a commander had or ought to have had some degree of knowledge that a certain action would cause a particular bad outcome.[39] Given the intrinsic complexity of the operation of software on LAWS, it would be very difficult to determine and prove the degree of knowledge the commander should possess and what degree of information available to the commander is of such a nature as to hold him or her responsible. Especially since the influence that the commander will have on the learning of subordinates will change drastically. In traditional situations, the commander trains the subordinates, whereas in the case of LAWS, the behavior will be largely determined by other actors than those who use it on the battlefield. Furthermore, it is uncertain to what extent the relationship between a human superior and human subordinate is analogous to that between a human superior and a non-human subordinate. Some authors emphasize that command responsibility rests on wrongdoings of subordinates and since LAWS cannot act consciously, it would be impossible to apply command responsibility analogously (Chengeta, 2016, 31).

6 Conclusion

The deployment of automated systems in tasks and contexts involving moral decision-making naturally raises ethical issues. The literature on LAWS, and by extension other autonomous decision systems, leading to responsibility gaps has grown rapidly in recent years. As should be clear by now, the literature does not provide a single answer to the questions of whether LAWS lead to responsibility gaps,

[38] For recent views and further explanations on predictability in LAWS, see among others: (Taddeo & Blanchard, 2022, p. 6–8; McFarland, 2020, p. 59–66; Holland Michel, 2020).

[39] According to an ordinary conception of responsibility attribution, it is only fitting to hold someone responsible if the agent can foresee that the device will or is likely to create a certain kind of outcome. This is usually termed the epistemic, or knowledge, condition and many philosophers agree that such a requirement is a necessary condition for moral responsibility. See among others: (Fischer & Tognazzini, 2009; Fischer & Ravizza, 2000, 13; Zimmerman, 1997).

whether it constitutes a new moral problem, and whether solutions can be found for it. I have attempted to increase understanding of the problem of responsibility gaps for LAWS by exposing some underlying premises. Furthermore, I have shown in this article that, if we accept the existence of the responsibility gap in a narrow sense, it is not so easy to simply close the gap. Moreover, I have pointed out the special nature of the technology and the fact that the divide between LAWS and military soldiers might be smaller than sometimes initially thought. A thorough solution to the problem of the responsibility gap would be one that both fully recognizes the problem and does not treat it as a mere empirical problem, and at the same time is able to reflect the increasingly autonomous nature of the technology without running the risk of anthropomorphizing LAWS or exempting all human actors involved from any responsibility.

Acknowledgements For helpful feedback and discussions, the author wishes to thank the reviewers of this paper, Carl Ceulemans, Lode Lauwaert, Stefan Rummens, Fabio Tollon, and Andrew Rebera.

Author Contribution Single-authored.

Funding This work was supported by the Belgian Defense – Royal Higher Institute of Defense (grant number HFM20-03).

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The author declares no competing interests.

References

Amoroso, D., & Giordano, B. (2019). Who Is to Blame for Autonomous Weapons Systems' Misdoings? In E. Carpanelli & N. Lazzerini (Eds.), *Use and Misuse of New Technologies* (pp. 211–232). Springer International Publishing. https://doi.org/10.1007/978-3-030-05648-3_11

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. https://doi.org/10.1017/S1816383112000768

Bo, M., Bruun, L., & Boulanin, V. (2022). *Retaining Human Responsibility in the Development and Use of Autonomous Weapon Systems: On Accountability for Violations of International Humanitarian Law Involving AWS*. Stockholm International Peace Research Institute. https://doi.org/10.55163/AHBC1664

Boulanin, V., & Verbruggem, M. (2017). *Mapping the development of autonomy in weapon systems*. 147.

Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (Vol. 8, pp. 63–74). John Benjamins Publishing Company. https://doi.org/10.1075/nlp.8.11bry

Calo, R. (2015). Robotics and the Lessons of Cyberlaw. *California Law Review, 103*, 513–563.

Champagne, M., & Tonkens, R. (2015). Bridging the Responsibility Gap in Automated Warfare. *Philosophy & Technology, 28*(1), 125–137. https://doi.org/10.1007/s13347-013-0138-3

Chengeta, T. (2016). Accountability gap: Autonomous weapon systems and modes of responsibility in international law. *Denver Journal of International Law and Policy*, *45*(1).

Crawford, N. C. (2007). Individual and Collective Moral Responsibility for Systemic Military Atrocity. *Journal of Political Philosophy, 15*(2), 187–212. https://doi.org/10.1111/j.1467-9760.2007.00278.x

Crawford, N. C. (2013). Organizational Responsibility. In *Accountability for Killing: Moral Responsibility for Collateral Damage in America's Post-9/11 Wars* (p. 92). Oxford University Press.

Crootof, R. (2016). War torts: Accountability for autonomous weapons. *University of Pennsylvania Law Review, 164*(6), 1347–1402.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology, 18*(4), 299–309. https://doi.org/10.1007/s10676-016-9403-3

Danaher, J. (2022). Tragic Choices and the Virtue of Techno-Responsibility Gaps. *Philosophy & Technology, 35*(2), 26. https://doi.org/10.1007/s13347-022-00519-1

de Jong, R. (2020). The Retribution-Gap and Responsibility-Loci Related to Robots and Automated Technologies: A Reply to Nyholm. *Science and Engineering Ethics, 26*(2), 727–735. https://doi.org/10.1007/s11948-019-00120-4

de Lima, T., & Royakkers, L. (2015). A Formalisation of Moral Responsibility and the Problem of Many Hands. In I. van de Poel, L. Royakkers, & S. Zwart, *Moral responsibility and the problem of many hands* (1 [edition], pp. 93–131). Routledge.

Laura A Dickinson. (2019). Lethal Autonomous Weapons Systems: The Overlooked Importance of Administrative Accountability. In R. T. P. Alcala & Eric Talbot Jensen (Eds.), *The Impact of Emerging Technologies on the Law of Armed Conflict* (p. 27). Oxford University Press.

Dinstein, Y. (2018). Autonomous Weapons and International Humanitarian Law. In W. Heintschel von Heinegg, R. Frau, & T. Singer (Eds.), *Dehumanization of Warfare* (pp. 15–20). Springer International Publishing. https://doi.org/10.1007/978-3-319-67266-3_2

Dremliuga, R. (2020). General Legal Limits of the Application of the Lethal Autonomous Weapons Systems within the Purview of International Humanitarian Law. *Journal of Politics and Law, 13*(2), 115. https://doi.org/10.5539/jpl.v13n2p115

Egeland, K. (2016). Lethal Autonomous Weapon Systems under International Humanitarian Law. *Nordic Journal of International Law, 85*(2), 89–118. https://doi.org/10.1163/15718107-08502001

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

Fischer, J. M., & Ravizza, M. (2000). *Responsibility and control: A theory of moral responsibility* (1st pbk. ed). Cambridge University Press.

Fischer, J. M., & Tognazzini, N. A. (2009). The Truth about Tracing. *Noûs, 43*(3), 531–556. https://doi.org/10.1111/j.1468-0068.2009.00717.x

Furlough, C., Stokes, T., & Gillan, D. J. (2021). Attributing Blame to Robots: I. The Influence of Robot Autonomy. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 63*(4), 592–602. https://doi.org/10.1177/0018720819880641

Galliott, J. (2020). No Hands or Many Hands? Deproblematizing the Case for Lethal Autonomous Weapons Systems. In A. E. Eckert & S. C. Roach (Eds.), *Moral responsibility in twenty-first-century warfare: Just war theory and the ethical challenges of autonomous weapons systems* (pp. 155–179). State University of New York.

Gunkel, D. J. (2020). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology, 22*(4), 307–320. https://doi.org/10.1007/s10676-017-9428-2

Gunkel, D. J. (2020a). Perspectives on Ethics of AI: Philosophy. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 537–553). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.35

Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and Information Technology, 15*(2), 99–107. https://doi.org/10.1007/s10676-012-9301-2

Himmelreich, J. (2019). Responsibility for Killer Robots. *Ethical Theory and Moral Practice, 22*(3), 731–747. https://doi.org/10.1007/s10677-019-10007-9

Holland Michel, A. (2020). *The Black Box, Unlocked: Predictability and Understandability in Military AI*. United Nations Institute for Disarmament Research. https://doi.org/10.37559/SecTec/20/AI1

Johnson, A. M., & Axinn, S. (2013). The morality of autonomous robots. *Journal of Military Ethics, 12*(2), 129–141. https://doi.org/10.1080/15027570.2013.818399

Kaiserman, A. (2021). Responsibility and the 'Pie Fallacy.' *Philosophical Studies, 178*(11), 3597–3616. https://doi.org/10.1007/s11098-021-01616-1

Köhler, S. (2020). Instrumental Robots. *Science and Engineering Ethics, 26*(6), 3121–3141. https://doi.org/10.1007/s11948-020-00259-5

Köhler, S., Roughley, N., & Sauer, H. (2017). Technologically blurred accountability? In C. Ulbert, P. Finkenbusch, E. Sondermann, & T. Debiel (Eds.), *Moral Agency and the Politics of Responsibility* (1st ed.). Routledge. https://doi.org/10.4324/9781315201399

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology, 24*(3), 36. https://doi.org/10.1007/s10676-022-09643-0

Lagioia, F., & Sartor, G. (2020). AI Systems Under Criminal Law: A Legal Analysis and a Regulatory Perspective. *Philosophy & Technology, 33*(3), 433–465. https://doi.org/10.1007/s13347-019-00362-x

List, C. (2021). Group Agency and Artificial Intelligence. *Philosophy & Technology, 34*(4), 1213–1242. https://doi.org/10.1007/s13347-021-00454-7

Margulies, P. (2019). Making autonomous weapons accountable: Command responsibility for computer-guided lethal force in armed conflicts. In J. D. Ohlin (Ed.), *Research handbook on remote warfare* (Paperback edition, pp. 405–442). Edward Elgar Publishing.

Marino, D., & Tamburrini, G. (2020). Learning robots and human responsibility. In W. Wallach & P. Asaro (Eds.), *Machine Ethics and Robot Ethics* (1st ed., pp. 377–382). Routledge. https://doi.org/10.4324/9781003074991-33

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McDougall, C. (2019). Autonomous weapon systems and accountability: Putting the cart before the horse. *Melbourne Journal of International Law, 20*, 58–87.

McFarland, T. (2020). *Autonomous weapon systems and the law of armed conflict: Compatibility with international humanitarian law*. Cambridge University Press.

Mettraux, G. (2009). The Resurgence of International Criminal Justice and the Rebirth of Command Responsibility. In *The Law of Command Responsibility* (pp. 13–20). Oxford University Press.

Nyholm, S. (2018). Attributing Agency to Automated Systems: Reflections on Human-Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics, 24*(4), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Ohlin, J. D. (2017). Remoteness and reciprocal risk. In *Research Handbook on Remote Warfare* (pp. 15–49). Edward Elgar Publishing.

Pagallo, U. (2013). Crimes. In U. Pagallo, *The Laws of Robots* (pp. 45–78). Springer Netherlands. https://doi.org/10.1007/978-94-007-6564-1_3

Robillard, M. (2018). No Such Thing as Killer Robots. *Journal of Applied Philosophy, 35*(4), 705–717. https://doi.org/10.1111/japp.12274

Roff, H. M. (2014). Killing in War: Responsibility, Liability, and Lethal Autonomous Robots. In *Routledge handbook of ethics and war: Just war theory in the twenty-first century* (Vol. 26, pp. 352–364). http://choicereviews.org/review/https://doi.org/10.5860/CHOICE.51-3176

Sander, B. (2010). Unravelling the Confusion Concerning Successor Superior Responsibility in the ICTY Jurisprudence. *Leiden Journal of International Law, 23*(1), 105–135. https://doi.org/10.1017/S0922156509990355

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*. https://doi.org/10.1007/s13347-021-00450-x

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. https://doi.org/10.3389/frobt.2018.00015

Santoro, M., Marino, D., & Tamburrini, G. (2008). Learning robots interacting with humans: From epistemic risk to responsibility. *AI & Society, 22*(3), 301–314. https://doi.org/10.1007/s00146-007-0155-9

Sassóli, M. (2014). Autonomous Weapons and International Humanitarian Law: Advantages, open technical questions and legal issues to be clarified. *International Law Studies, 90*, 308–340.

Sassòli, M., & Nagler, P. (2019). *International humanitarian law: Rules, controversies, and solutions to problems arising in warfare*. Edward Elgar Publishing.

Saxon, D. (2016). Autonomous Drones and Individual Criminal Responsibility. In E. Di Nucci & F. S. de Sio (Eds.), *Drones and Responsibility: Legal, Philosophical, and Sociotechnical Perspectives on Remotely Controlled Weapons* (1st ed., pp. 17–46). Routledge. https://doi.org/10.4324/9781315578187

Schmitt, M. N. (2012). Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics. *SSRN Electronic Journal*, 1–37.https://doi.org/10.2139/ssrn.2184826

Schulzke, M. (2013). Autonomous Weapons and Distributed Responsibility. *Philosophy & Technology, 26*(2), 203–219. https://doi.org/10.1007/s13347-012-0089-0

Schwarz, E. (2021). Autonomous Weapons Systems, Artificial Intelligence, and the Problem of Meaningful Human Control. *Philosophical Journal of Conflict and Violence, 5*(1), 53–72. https://doi.org/10.22618/TP.PJCV.20215.1.139004

Seher, G. (2016). Intelligente Agenten als „Personen" im Strafrecht? In S. Gless & K. Seelmann (Eds.), *Intelligente Agenten und das Recht* (pp. 45–60). Nomos Verlagsgesellschaft mbH & Co. KG. https://doi.org/10.5771/9783845280066-45

Simmler, M., & Markwalder, N. (2019). Guilty Robots? – Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence. *Criminal Law Forum, 30*(1), 1–31. https://doi.org/10.1007/s10609-018-9360-0

Simpson, T. W., & Müller, V. C. (2016). Just War and Robots' Killings. *The Philosophical Quarterly, 66*(263), 302–322. https://doi.org/10.1093/pq/pqv075

Smith, P. T. (2019). Just research into killer robots. *Ethics and Information Technology, 21*, 281–293.

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy, 24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Taddeo, M., & Blanchard, A. (2022). Accepting Moral Responsibility for the Actions of Autonomous Weapons Systems—A Moral Gambit. *Philosophy & Technology, 35*(3), 78. https://doi.org/10.1007/s13347-022-00571-x

Taylor, I. (2021). Who Is Responsible for Killer Robots? Autonomous Weapons, Group Agency, and the Military-Industrial Complex. *Journal of Applied Philosophy, 38*(2), 320–334. https://doi.org/10.1111/japp.12469

The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches. (2017). United Nations Institute for Disarmament Research. https://www.unidir.org/files/publications/pdfs/the-weaponization-of-increasingly-autonomous-technologies-concerns-characteristics-and-definitional-approaches-en-689.pdf

Tigard, D. W. (2020). There Is No Techno-Responsibility Gap. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00414-7

Tigard, D. W. (2021). Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible. *Cambridge Quarterly of Healthcare Ethics, 30*(3), 435–447. https://doi.org/10.1017/S0963180120000985

Tollon, F. (2022). Is AI a Problem for Forward Looking Moral Responsibility? The Problem Followed by a Solution. In E. Jembere, A. J. Gerber, S. Viriri, & A. Pillay (Eds.), *Artificial Intelligence Research* (Vol. 1551, pp. 307–318). Springer International Publishing. https://doi.org/10.1007/978-3-030-95070-5_20

van de Poel, I. (2015). *Moral responsibility and the problem of many hands (1 [edition])*. Routledge.

van de Poel, I., Nihlén Fahlquist, J., Doorn, N., Zwart, S., & Royakkers, L. (2012). The Problem of Many Hands: Climate Change as an Example. *Science and Engineering Ethics, 18*(1), 49–67. https://doi.org/10.1007/s11948-011-9276-0

Van Severen, S., & Vander Maelen, C. (2021). Killer robots: Lethal autonomous weapons and international law. In J. de Bruyne & C. Vanleenhove (Eds.), *Artificial intelligence and the law* (pp. 151–172). Intersentia.

Zimmerman, M. J. (1985). Sharing Responsibility. *American Philosophical Quarterly, 22*(2), 115–122.

Zimmerman, M. J. (1997). Moral Responsibility and Ignorance. *Ethics, 107*(3), 410–426. https://doi.org/10.1086/233742

Zimmerman, M. J. (2015). Varieties of Moral Responsibility. In R. Clarke, M. McKenna, & A. M. Smith (Eds.), *The Nature of Moral Responsibility* (pp. 45–64). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199998074.003.0003

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