AI and Ethics (2025) 5:1757–1767
https://doi.org/10.1007/s43681-024-00512-8

ORIGINAL RESEARCH

Command responsibility in military AI contexts: balancing theory and practicality

Ann-Katrien Oimann[1,2] · Adriana Salatino[3]

Received: 8 April 2024 / Accepted: 22 June 2024 / Published online: 22 July 2024

Abstract
Artificial intelligence (AI) has found extensive applications to varying degrees across diverse domains, including the possibility of using it within military contexts for making decisions that can have moral consequences. A recurring challenge in this area concerns the allocation of moral responsibility in the case of negative AI-induced outcomes. Some scholars posit the existence of an insurmountable "responsibility gap", wherein neither the AI system nor the human agents involved can or should be held responsible. Conversely, other scholars dispute the presence of such gaps or propose potential solutions. One solution that frequently emerges in the literature on AI ethics is the concept of command responsibility, wherein human agents may be held responsible because they perform a supervisory role over the (subordinate) AI. In the article we examine the compatibility of command responsibility in light of recent empirical studies and psychological evidence, aiming to anchor discussions in empirical realities rather than relying exclusively on normative arguments. Our argument can be succinctly summarized as follows: (1) while the theoretical foundation of command responsibility appears robust (2) its practical implementation raises significant concerns, (3) yet these concerns alone should not entirely preclude its application (4) they underscore the importance of considering and integrating empirical evidence into ethical discussions.

Keywords Command responsibility · Responsibility gap · Artificial intelligence · AI ethics · Psychology

1 Introduction

AI is presently being used in virtually every domain, including in the military. There are currently many AI applications in the military context and their adoption is accelerating, including those requiring ethical evaluation and judgement. One application that is receiving a lot of attention is (assisted) decision making systems for targeting and engagement purposes[1], particularly lethal autonomous weapon systems (LAWS), and questions arise about how such systems should be regulated and how International Humanitarian Law (IHL) should deal with them [1–5]. Alongside the focus on IHL, another critical issue is the allocation of responsibility for actions taken by these systems. This concern is prominent in both international debates addressing the governance of military AI and academic discussions. A significant concern in ethical debates revolves around the attribution of responsibility for these systems in the case something goes wrong, commonly referred to as 'responsibility gaps'. This term, introduced by Andreas Matthias in the context of autonomous machines (2004) and later applied to LAWS by Robert Sparrow [6], continues to be a focal point of discussion in both philosophical and legal realms [7–9]. In the context of LAWS, responsibility

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

Adriana Salatino
Adriana.salatino@mil.be

[1] Department of Behavioural Sciences, Royal Military Academy, Hobbema 8, Brussels 1000, Belgium

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

[3] Department of Life Sciences, Royal Military Academy, Hobbema 8, Brussels 1000, Belgium

[1] See in this regard for example the recent use of AI-systems 'the Gospel' and 'Lavender' that were used for assisting with target selection by the Israel Defense Forces in Gaza: https://www.theguardian.com/world/2024/apr/03/israel-gaza-ai-database-hamas-airstrikes?CMP=Share_AndroidApp_Other and https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets.

gaps manifest when it seems appropriate to hold someone responsible for a certain (bad) outcome, but according to our standard theories of moral responsibility attribution, there is no suitable target of blame.

Although it is important to recognize that neither autonomous systems in general nor LAWS in particular are simply machines but rather sociotechnical systems involving both machine (hardware and software) components and a multitude of human actors (developers, operators, users) leading some authors to state that the responsibility gap is a many hands problem [10], the predominant focus in discussions responsibility gaps and LAWS has been on assigning backward-looking responsibility. This debate revolves around the argument that due to the growing autonomy of the system - characterized by its self-learning ability and ability to adjust behavior in response to feedback from the environment - neither the machine nor the humans involved can be held responsible. The machine, although potentially causally responsible, lacks moral agency, which means that, according to standard theory, it cannot be held responsible. At the same time, humans are considered to lack the necessary control and/or knowledge, essential conditions for assigning moral responsibility. It would be therefore unfair to hold either party responsible [6, 8]. In an effort to address this so-called responsibility gap, multiple solutions have been proposed. Among them are authors advocating for a revision of the conditions described in our standard theories of responsibility attribution. By modifying the standard conditions, it becomes feasible to hold humans (or machines) responsible.[2] Proposals in this category include to conceptually engineer the concept of responsibility differently [16]; the introduction of a 'blank check' liability, where humans hold themselves responsible for the actions of military robots [17], and legal proposals suggesting the adoption of strict liability in criminal law.

Responsibility attribution is generally based on the effects directly caused by an action. This becomes challenging in the context of LAWS due to their self-learning and autonomous capabilities. Multiple authors have argued against direct individual criminal liability for operators or military commanders in the context of LAWS, emphasizing that such liability requires intent or recklessness and a direct causal link between action and outcome [18]; Dickinson, 2019; Chengeta, [19] pp. 16–27; Crootof, [7] p. 1376; Egeland, [3] p. 106; Saxon, [20] p. 28). Given that neither operators nor commanders are directly involved in the execution of attacks by LAWS and lack a guilty mind due to the very limited meaningful human control, establishing direct individual criminal liability appears highly unlikely. However, an intriguing solution proposed to address responsibility gaps in LAWS is rooted in responsibility ascription within military hierarchies, specifically through the concept of command responsibility.[3] The doctrine of command responsibility is part of customary international law, as codified in art. 28 of the ICC Statute, art. 86–87 Additional Protocol I to the Geneva Conventions and art. 7 (3) of the ICTY Statute and interpreted in case law of the UN international tribunals for the former Yugoslavia and for Rwanda (among others in the Delalić case, Blaskić case, Kayishema and Ruzindana case). The doctrine holds military leaders criminally responsible for crimes committed by their subordinates, constituting a form of liability for omissions related to the acts of subordinates rather than a separate criminal offense [21], p. 18). This approach is interesting because it enables commanders to be held responsible for indirectly caused effects. Responsibility can be attributed based on the causal-like relationship (supervision) in which humans stand, allowing responsibility to follow the lines of supervision.

The solution of using command responsibility to solve responsibility gaps in LAWS has been both proposed and criticized in AI ethics and legal literature.[4] However, what has been lacking in such ethical and legal evaluations and remains under researched is the link with recent empirical psychological studies. This integration could not only lead to a better understanding of the practical implications of applying theoretical frameworks such as the doctrine of command responsibility in real-world contexts, but also advance the current ongoing theoretical debate, as empirical studies can serve to validate or challenge theoretical assumptions in the literature and identify new patterns and trends in the assignment of responsibility. Therefore, this article specifically addresses responsibility gaps within the context of LAWS and aims to examine the extent to which applying the doctrine of command responsibility as a solution to these gaps creates a moral gap. What is presented morally as well-argued and persuasive is not necessarily consistent with how this moral issue is perceived from a psychological perspective. The purpose is to anchor discussions in empirical realities rather than relying exclusively

[2] For clarity and completeness, this type of solution theoretically allows advocating for the responsibility of the system itself. However, so far, all existing law is predicated on human responsibility and accountability, and even in AI ethics, there is currently no serious debate about holding AI systems or LAWS themselves responsible [11]. Some exceptions who do discuss the possibility of attributing moral responsibility to systems themselves are: [12–15].

[3] The generic expression "superior responsibility" is also often used since the doctrine also applies to civilian leaders. However, they are often used interchangeably and in terms of content, there is no substantial difference between the two expressions. In the remainder of this article, we will use the term 'command responsibility' because the article is set in a military context.

[4] Authors who believe that the problem of assigning responsibility can be (at least partly) solved by looking at the hierarchical structure in the military include: Schulzke [22–24], Schmitt [25]. For authors who don't believe in the feasibility, see: [7, 8, 19, 26, 27].

on normative arguments. The ultimate goal of this paper is to assess the extent to which the analogous application of command responsibility to LAWS is feasible and desirable.

In Sect. 1, we explore the initial theoretical plausibility of applying the doctrine of command responsibility to autonomous systems, considering factors such as anthropomorphic tendencies, perceptions of responsibility of commanders when deploying autonomous systems, and the impact of interaction with autonomous systems on moral decision-making and Sense of Agency (SoA). We suggest that commanders may be suitable candidates for responsibility. In Sect. 2 we will argue that it may not be prudent to apply the doctrine analogous based on research in traditional hierarchical settings, which suggest that neither coerced agents nor commanders experience agency over their actions, potentially leading to a genuine responsibility gap. However, in Sect. 3 we argue that this gap may not inherently pose a problem since not all normative solutions need to be grounded in descriptive facts[5], drawing parallels with traditional hierarchical settings where we generally attribute responsibility to commanders despite their apparent lack of sense of responsibility. Section 4 concludes by advocating caution in applying the doctrine to non-human agents and gives practical and theoretical reflections to demonstrate why the gap between empirical facts and ethical solutions should not be too wide.

2 Viability of applying command responsibility to artificial subordinates

In this section, we explore the viability of applying the doctrine of command responsibility - traditionally applicable to situations involving human superiors and human subordinates [$C^H \wedge S^H$] - to scenarios that include an artificial subordinate [$C^H \wedge S^A$]. The doctrine is a jurisprudential doctrine in international criminal law that aims to hold military commanders accountable for war crimes committed by their subordinates. The doctrine has been developed and applied in varied ways by ad hoc tribunals and the International Criminal Court (ICC) leading to inconsistent codification in international agreements [28], pp. 265–266). Generally, it holds superiors liable if they have actual control over a subordinate, know or have reason to know of the subordinate's criminal acts, and fail to take necessary and reasonable measures to prevent or punish them [7], p. 1378). Ethically, the application may seem plausible, as commanders are generally considered responsible for adverse outcomes caused by human subordinates, despite these subordinates often having significant discretionary power. Moreover, the argument that the concept of command responsibility was initially formulated to regulate the interactions between humans on the battlefield is no prima facie reason for excluding its adaptation to address new challenges involving artificial entities. The application of the doctrine as a possible answer to the responsibility gap is not new. It has been defended in the AI ethics and legal literature by Himmelreich [22–24] among others in the context of LAWS. In this section, we argue that such an adaptation might not only make sense from an ethical or legal standpoint but also from an empirical perspective. We outline three reasons drawn from recent literature in psychology, each of which we will discuss in turn.

The first reason for considering AI as a subordinate agent is our human tendency for anthropomorphizing. Anthropomorphizing refers to the inclination to attribute distinctively human characteristics to nonhuman entities [29]. Peter Singer and Joel Garreau [30] have reported already more than a decade ago the intriguing tendency of humans to form relationships with machines. They illustrated this phenomenon by examining the behavior of US soldiers in Iraq and Afghanistan, who developed unexpectedly close personal bonds with their PackBots by giving them names, awarding them battlefield promotions, risking their lives to protect the 'life' of the robot and mourning their 'deaths' [30, 31]. This unique human-machine bonding is a result of the system's integration within the military unit and the role that it plays in battlefield operations. Another factor is the shape of the machine and perceptions of similarity to humans, as recent research indicates a favorable effect of anthropomorphic design features on human-related outcomes [32]. A further instance of this tendency affecting people's judgments and decision-making includes a greater reluctance of humans to sacrifice machines [33]. Thus, viewing AI as a subordinate agent is not as far-fetched as one may think, and is plausible from that viewpoint. Moreover, a growing body of psychological studies in recent years has shown that we tend to attribute moral responsibility to nonhuman agents, leading us to be willing to blame robots [34–46]. This tendency is particularly noticeable in scenarios where the robot is described as autonomous compared to scenarios where the robot is described as nonautonomous, suggesting that the degree to which people view them as social actors and attribute blame to them depends on the perceived degree of autonomy [47]. Worth mentioning is a recent paper by Kneer and Christen [38] p. 3), who conducted a cross-cultural empirical study using Robert Sparrow's famous example of an autonomous weapon system committing a war crime[6] among Japanese, German and U.S. participants. The study concluded that people show a considerable willingness to hold autonomous weapon systems morally responsible. This finding seems to

[5] It can be argued that it remains uncertain whether any normative solutions require such grounding, we come back to this fact in Sect. 3.

[6] See: [6].

contradict the hypothesis often put forward in philosophy literature that people find morally responsible machines absurd and demonstrates that people are far from dismissive of the possibility of assigning moral responsibility to a machine.

A second reason for considering the extension of the concept of command responsibility to non-human subordinates is the notion that a commander is held responsible not only for the war crimes committed by a human pilot, but also when dispatching an autonomous system. This follows from the experiment conducted by Kneer and Christen [38], in which it was clearly demonstrated that commanders were deemed equally responsible in both conditions (Japan), and even significantly *more* responsible when dispatching an autonomous system, in contrast to situations where human pilots were deployed (Germany and the US). This is consistent with the findings by Caspar et al. [48] indicating that explicit responsibility self-ratings were higher when the commander gave orders to a robot agent compared to when the commander gave orders to a human agent (p. 17). One possible explanation for the different levels of responsibility attribution could be that in the traditional case [$C^H \wedge S^H$], subordinates are believed to have more discretionary power compared to situations involving autonomous systems [$C^H \wedge S^A$]. Consequently, the autonomy of the subordinates influences the perception of responsibility of the commander. This may be attributed to the belief that commanders have a more active and direct role in the planning and deployment of military operations in these cases. Whether this influence is truly greater in [$C^H \wedge S^A$] situations would need to be assessed on a case-by-case basis, depending on the capabilities of the system. Nevertheless, Caspar and colleagues' experiment in their 2021 paper appears to capture the intuition that commanders in these cases still determine the general parameters under which the systems operate, such as where, when, how and against whom military force may be used. Machines do not create tasks ex nihilo and are always restrained by hierarchical orders. In essence, the experiment supports the proposal by philosophers and lawyers that human agents can be held responsible for adverse outcomes caused by machines based on their supervisory role.

Thirdly, moral decision-making increases the SoA [49], and conscious engagement of people in morally challenging tasks does not seem to be adversely affected by the interaction with an autonomous system [50]. In preparation for explaining this, let us first get clarity on what SoA is. SoA refers to the awareness that humans have of being the authors of their actions and thus of the consequences of these actions [51–53]; Pyasik, Salatino et al., [54]). SoA enables us to perceive ourselves as causal agents [55] and is thus a precursor of feeling responsible for a deed. It is recognized as an important aspect of human consciousness, and it is closely related to moral responsibility [49]; Caspar, Christensen, Cleeremans, & Haggard, [56]). However, while responsibility is an explicit and social concept, SoA is often measured implicitly, among other technique with the *Intentional Binding* (IB) effect. The IB refers to the subjective compression of the interval between an action and its outcome observed in active, but not passive, movement: participants are asked to estimate the time interval between an action they perform and its consequences [57–59]. A series of previous studies have shown that the time estimation between action and outcome is a valid implicit, quantitative measure of SoA [60–62] and is preferable to a subjective measurement of responsibility, which is subject to social desirability and other biases, such as the self-serving bias (e.g., Blackwood et al., [63, 64]. This finding implies that a commander might be a suitable candidate to be held responsible, as the engagement and sense of agency remain intact despite the involvement of autonomous systems [50].

In summary, when we combine the three aforementioned reasons, a preliminary compelling argument emerges. Humans display a clear tendency for anthropomorphism, even exhibiting a willingness to hold robots responsible. Considering this, the S^A-part of the hypothesis seems feasible. Moreover, commanders are held equally or even more responsible than human subordinates. Additionally, the fact that people engaged in morally challenging tasks are not negatively affected by interactions with autonomous systems suggests the possibility that a commander might be a suitable candidate for being held morally responsible.

3 Examining concerns in practical application and hierarchical dynamics

However, relying solely on the results of Sect. 1 is not a robust argument for extending responsibility. This is not only true from a logical reasoning standpoint but is further underscored by recent empirical research on the influence of hierarchical relationships on the attribution of moral responsibility. In this section, we argue that such an extension might not be prudent due to a potential risk of a decrease in the SoA of the commander.

Research conducted in traditional [$C^H \wedge S^H$] relationships indicates that the SoA of a commander decreases in hierarchical settings, as well as the SoA of the subordinate, so that both the commander and the subordinate feel less responsible in hierarchical settings [48, 65]. In the 2018 study, Caspar and colleagues investigated, in a commander-subordinate relationship, whether the SoA and responsibility pass from the person who receives orders to the person who gives them. In the experiment,

volunteers took turns to play the roles of 'commander', 'agent' or 'victim' in a task where the commander instructed the agent to deliver painful shocks to the 'victim'. They tested the implicit sense of agency but also explicitly questioned responsibility. The results showed that the SoA decreased when agents (i.e. subordinates) received orders, compared to when they freely chose which action to execute, and that SoA decreased in commanders when they commanded agents to administer the shock on their behalf, compared to when they acted on their own. In other words, the results suggest that coercive situations potentially undermine the sense of responsibility of agents and commanders, as both experience less agency over their actions and its consequences, compared to a situation in which they would act on their own.

In the 2021 study, the methodology was roughly the same. In addition, here, neuroimaging techniques (i.e., functional magnetic resonance imaging-fMRI and electroencephalography-EEG) were employed to investigate the neuronal activity (involved in SoA and in empathy for pain) in hierarchical situations. One of the interesting findings there was that a difference in brain activation could be observed between participants who could freely decide which orders to give to another agent ("free commanders") and participants who were free to decide to execute the orders ("free agents"). The brain activity in the relevant areas linked with empathy and emotional social perception was higher among the "free agents" compared to the "free commanders" suggesting that actually performing the action is important for social cognition and that despite commanders having the same decisional power, being further away from the outcome of that action does have an influence [48], p. 14). This is consistent with the findings of comparing activation between subordinates giving orders, where it was discovered that the subordinate agent had higher brain activation in empathy related areas compared to the commanders. This again suggests that acting has a higher influence in empathic response than having decisional power [48], p. 26).

When we consider the prospect of resolving responsibility gaps through the mechanism of command responsibility and combine it with the preliminary conclusions from Sect. 1 — specifically, that it is plausible to treat the a-moral agent as a (non-human) subordinate — we might find ourselves in a situation where there is also a decrease in the SoA of the commander, similar to the traditional $[C^H \wedge S^H]$ situations. This echoes the findings of Caspar et al., indicating that the true responsibility gap emerges not solely from an inability to blame the machine or the human, but is rooted in hierarchical relationships. This holds true not only in human-machine situations but also in human-human situations.

In this regard, it is helpful to return to the research conducted by Kneer and Christen [38], who specifically investigated hierarchical situations involving autonomous weapons (a-moral subordinates). One noteworthy finding was that a commander dispatching a robot pilot was consistently deemed significantly *less* responsible for the harm than a human pilot in traditional situations. While the commander in these scenarios was generally considered equal or more responsible than in situations with a human pilot, the attributed level of responsibility to the commander was still *less* compared to the human pilot. This indicates that, in the eye of the person judging the situation, not all responsibility is fully transferred from the robot pilot to the commander. These findings are consistent with the research of Bigman et al. [66] and Shank, DeSanti, et al. [67] revealing that blame attribution to AI for moral wrongdoing is less than that for humans, and humans monitoring AI are faulted less than those working in teams composed solely of humans.

These findings delve into the core of the problem, providing empirical evidence for what philosophers have theorized. In human delegation, an authoritative agent (commander) delegates tasks or competences to a subordinate with a portion of responsibility, but retains a variable portion of responsibility. When bad consequences arise from the actions of subordinates, not only are the subordinates held responsible, but a share of the responsibility is shifted to the superior if certain conditions are met. Applying this framework to situations where commanders delegate tasks or competences to amoral subordinates, such as autonomous machines $[C^H \wedge S^A]$, seems similar and there is no apparent reason why part of the responsibility cannot be shifted to the superiors in cases of negative consequences caused by these amoral subordinates. However, a crucial distinction emerges: in human-to-human delegation, it remains possible to hold the subordinate responsible for the underlying actions leading to adverse outcomes. Conversely, in delegations to amoral agents, it becomes (normatively) impossible to assign responsibility to the subordinates for their actions. When we transfer responsibility for the actions of amoral subordinates to commanders, the distribution of this responsibility becomes unclear. Questions arise regarding the nature of the commander's responsibility and what responsibilities, if any, remain with the subordinates after the transfer. A complete transfer of responsibility from machine to user appears unfeasible due to a distinction between two types of responsibility: (1) moral outcome responsibility, which pertains to the responsibility for the actual consequences or outcomes of the

actions taken by autonomous machines, and (2) moral responsibility for the use or deployment, which relates to the user's responsibility for utilizing a machine with the potential to cause unfortunate results. This distinction highlights the challenge of 'residual responsibility'– the portion of (1) moral outcome responsibility that has not been fully transferred from the subordinates to the commander. To address the responsibility gap as the pessimists[7] conceive it, a complete transfer of (1) would be required. However, it seems to be that commanders are responsible for (2). Consequently, a comprehensive solution to the responsibility gap still appears elusive and the gap, as pessimists envision it, seems to be inherently unsolvable because the 'remaining' responsibility can never be fully transferred. Nevertheless, it is worth mentioning in this context Frank Hindriks and Herman Veluwenkamp, who argue that this kind of reasoning is based on a wrong assumption, namely that the amount of blame that is appropriate in the nearby possible world is an appropriate reference point for the actual world, but that there is no good reason to think that this assumption is correct. Therefore, they argue that this 'deficit conception' of responsibility gaps is problematic and that there are no responsibility gaps [71], pp. 5–6).

Kneer and Christen [38] did not distinguish between the two types of responsibility in their experiments. However, they also ended up with a 'remaining' share of responsibility not transferred to the commanders. If all responsibility had been fully transferred from the (amoral) subordinate to the commander, their results would have shown an equal amount of responsibility allocated to the commander dispatching a robot pilot compared to the human pilot (subordinate). In summary, this not only highlights that empirically speaking the true responsibility gap seems to exist in every hierarchical relationship, regardless of being a human-human or a human-machine relationship, but also suggests that human-machine situations leave us with an untransferred share of responsibility.

4 Navigating the interplay between ethics and descriptive realities

Not all normative solutions necessarily need to be grounded in purely descriptive facts. In fact, this is rarely the case, and deriving a moral 'ought' from an 'is' (action) is a fallacy.[8] The questions we address in general in moral philosophy in relation to responsibility, differ substantially from those investigated in psychology. While moral philosophy attempts to answer 'who (if anyone) can be rightly held responsible for harm?', psychology explores the real human tendencies regarding assigning responsibility in such contexts. This exploration includes two distinct sub-questions, namely (1) 'what are people's retributivist moral dispositions?' (outsider-view) and (2) 'how responsible do people who are being held responsible actually feel?' (insider-view). Both implicit measurement methods, such as the IB, which allow measuring quantitatively the SoA, and explicit methods, like asking questions or self-reports, can be used in answering these questions. However, this does not mean that both methods are equally suited to answer both questions. For example, both methods are used to measure SoA, but the first question, finding out what the retributive moral dispositions of outsiders are, seems to be answered only by obtaining a direct report on how they attribute responsibility. Preliminary outcomes from psychology research relevant to command responsibility suggest that (1) people might not tend to (fully) blame commanders for bad outcomes caused by autonomous machines, and (2) commanders themselves feel less responsible for tasks carried out by subordinates. However, these descriptive facts are not reasons to believe that commanders cannot be held responsible or that people should not hold them responsible.

Therefore, the risk described above in section two needs not in principle be a problem for using the solution of command responsibility for negative outcomes caused by AI-subordinates. The reason for this is that the responsibility gap that would arise in such a scenario is not inherently more problematic than in traditional configurations [$C^H \wedge S^H$], where we generally attribute responsibility to commanders despite their apparent lack of felt responsibility. Especially in military contexts it seems plausible that commanders do not feel responsible for adverse outcomes caused by human subordinates, yet we do hold them responsible. This is done to ensure sufficient supervision by the commander, and this intention also holds true for the AI subordinate. One example worth mentioning in this regard is the Yamashita case. General Tomoyuki Yamashita was held ultimately responsible for numerous war crimes relating to the Manila massacre and many atrocities in the Philippines against unarmed civilians and prisoners of war between 9 October 1944 and 2 September 1945, and was sentenced to death, despite the

[7] We use the term 'pessimists' for authors who believe in the existence of responsibility gaps and believe it is an unsolvable problem such as de Jong [6, 8, 68, 69]. Santoni de Sio and Mecacci use the term 'fatalists' in this regard [70].

[8] This is known as the 'is-ought problem' that was described by David Hume [72]. It should be noted that there is also literature that denies the existence of this fallacy.

controversy surrounding the case.[9] It was argued that he was responsible for his subordinates and failed to maintain his duty as commander to control the operations and the members of his command, thereby permitting them to commit brutal atrocities.

The case was controversial because no evidence was presented to show that Yamashita had ordered the violence or that he had known about the acts, and a standard of liability set out by the military commission amounted to an objective form of liability pursuant to which a commander could be held criminally responsible for crimes committed by his troops despite the absence of control of the criminal acts of his subordinates and regardless of any awareness or knowledge on his part that such crimes had been committed [21]. Yamashita denied until the very end that there was a way for him to control all actions by all his subordinates because of a disruption in communication and command structure and that he had knowledge of the crimes committed by his subordinates. He claimed that he would have harshly punished them if he would have had that knowledge. So, in general, when we make moral and legal decisions, we do that based on normative reasoning that involves requirements. In the case of the application of command responsibility these include a relationship between commander and subordinates, the control of the commander over the actions of the subordinates and the knowledge that the commander has or should have had that a certain outcome is going to happen. As such, the reason the case was controversial was not because Yamashita himself didn't feel responsible or that it was proved that public outsiders didn't have a tendency to blame him, but the fact that the requirements that we generally use for attributing blame were seriously contested in this case.

In conclusion, many ethical solutions may not align with descriptive facts, and not all normative solutions need to be grounded in descriptive facts. It is incorrect to absolve people of responsibility solely because they don't feel responsible, just as it is incorrect to argue that people feel responsible and should therefore be held responsible. So where does this leave us in in the context of applying the doctrine of command responsibility to LAWS? Based on the empirical research, the theoretical solution of command responsibility becomes more plausible. However, empirically speaking there is a serious risk of encountering impediments to assigning responsibility to AI systems, similar to traditional hierarchical situations. It would be misguided to assert normatively that commanders are responsible purely based on descriptive factors such as people (not) holding commanders responsible or commanders (not) holding themselves responsible.

4.1 Recognizing the vital role of empirical evidence in ethical frameworks

Nevertheless, caution is warranted, and a degree of reluctance to decouple ethical solutions from empirical psychological studies is advisable because of the potential unwanted risks that may arise when our ethical solutions deviate significantly from descriptive empirical facts, including psychological studies. It is desirable that the gap between the two is not too broad, as thoughtful research benefits from some level of empirical support. Furthermore, it can be pointed out that it should at least prompt ethicists to try to understand why there is such a significant gap. In this regard three arguments can be raised in favor of why they should not diverge significantly. We will discuss them each in turn.

First, for risk mitigation, descriptive psychological facts can highlight potential unintended consequences of ethical solutions. Ensuring some alignment between ethical solutions and descriptive psychological facts ensures that ethical solutions are designed to minimize negative psychological impacts and behavioral side effects. An example of this, in light of our discussion on using the doctrine of command responsibility to address the problem of the responsibility gap in the context of non-human agents, is the risk that superiors may (un)intentionally try to evade responsibility more readily when delegating tasks to machines compared to humans. This possibility introduces the concern of a false diversion of moral responsibility [73]. By categorizing artificial agents as subordinates under the doctrine of command responsibility, there may be an unintended incentive for strategic moral scapegoating. This phenomenon involves successfully shifting punishment to the artificial agent, providing superiors with a means to evade punishment more effectively when tasks are delegated to machines rather than humans. Such a scenario raises ethical concerns, as it could lead to an overuse of machines or human-machine teaming in certain situations. This overreliance, driven by a potential escape route for superiors from responsibility, may compromise decision-making processes and ethical considerations when using autonomous systems. In essence, while the theoretical underpinnings of applying command responsibility to non-human agents may appear plausible, the potential for unintended consequences and strategic moral scapegoating, as demonstrated by empirical research, necessitates a careful and thorough evaluation of its practical implications. Balancing the need for accountability with the risk of misuse and overreliance

[9] It gave rise to two strident dissenting opinions by two of the members in the US Supreme Court.

on machines is crucial in determining the feasibility and desirability of extending the doctrine in the context of artificial intelligence and autonomous systems.

Second, we can point to practical concerns. The convergence of ethical solutions with descriptive psychological facts and empirical studies strengthens the foundation of ethical frameworks, ensuring that the proposed ethical solutions are feasible and making them more applicable, acceptable, and effective in real-world contexts. This also has benefits related to informed policymaking. Decision-makers, whether organizations or governments, benefit from ethical solutions that are aligned with empirical studies, as this helps to make informed decisions that are based on realistic understanding of human behavior. In other words, to increase public trust, ethical solutions that align with psychological facts avoid idealization and acknowledge the complexities and nuances of human behavior, steering clear of overly optimistic or pessimistic assumptions that may not reflect reality. It is also important to mention that by including empirical evidence, ethical solutions acknowledge the diversity of human experiences and behaviors, promoting an ethical framework that is relevant and considerate of different human realities. If we argue for a normative theory, such as applying the command responsibility framework and blaming human commanders —in other words, imposing moral responsibility on people— we better make sure that it works and that we model and cultivate the intuitions along these lines.

Third, there is a more theoretical consideration for why the gap should not be too broad. Ethics doesn't fall from the sky; it is developed by humans to regulate relationships. Our ethical theories aim to balance what people think and their intuitions and empirical findings on the one hand with the normative case on the other hand. The key is to strike this balance in a good way, by not solely relying on practical feasibility or falling into the trap that empirical research has the last word in everything but also by not ignoring empirical research and the reality of multiple people. Our normative theories don't come out of thin air, and so command responsibility is already influenced by our intuitions. The concept of moral responsibility appears to be intimately linked to and shaped by the fact that we are social human beings existing within a specific moral community. This can be illustrated in the case of command responsibility. One of the reasons for the development of the doctrine of command responsibility after the Second World War, where thousands of individuals were held accountable for their actions, was the realization that came that international criminal law was underdeveloped, and a deep chasm existed between what was generally regarded as morally repugnant and the range of conducts that qualified as criminal offenses under international law [21]. There was, therefore, a general awareness of the necessity for international law to catch up with basic moral standards. Furthermore, from a theoretical standpoint, looking at empirical research is important to ensure the flexibility and adaptability of ethical solutions. As our understanding of human behavior evolves, ethical frameworks that are informed by empirical studies can evolve based on new insights. Ethical solutions should be adaptive and open to incorporating new knowledge. For example, if empirical studies reveal non-intuitive aspects of human behavior, ethical frameworks may need to evolve to accommodate these insights. These non-intuitive empirical results may prompt deeper ethical reflection and can stimulate ethical discourse, as ethicists may need to critically examine the reasons behind certain intuitions and consider whether they align with broader ethical principles or biases.[10] Ethical frameworks are not static. If empirical studies challenge intuitions, it signifies that ethical frameworks should undergo an iterative process, continually refining and adapting to emerging knowledge.

5 Conclusion

In this paper, we have closely examined the concept of command responsibility, which is often proposed as a solution to address responsibility gaps emerging in systems with increasing autonomous capabilities, particularly within the context of LAWS. By analyzing recent psychological research, we have asserted that, initially, the solution appears theoretically plausible and well supported based on preliminary empirical evidence. However, upon deeper exploration of existing empirical studies on hierarchical human relationships, we identified significant obstacles in assigning responsibility to AI systems, paralleling the challenges in traditional hierarchical settings. Nevertheless, we argue that these empirical concerns alone should not dismiss the application of command responsibility to non-human entities. Instead, we emphasize the importance of integrating empirical realities into normative discussions and solutions. This not only serves practical purposes but also enhances the ethical discourse itself. By acknowledging and incorporating empirical evidence, we can refine our ethical frameworks and develop more robust solutions to address emerging challenges in AI governance within military contexts.

Acknowledgements A previous version of this paper was presented at an Ethics Symposium in November 2023, organized by prof. Deane-Peter Baker and supported by the School of Humanities and Social Sciences at the UNSW Canberra. For comments that greatly improved the quality of this article, we would like to sincerely thank the organizers

[10] See in this regard also: [74].

and participants of this event. Furthermore, we gratefully acknowledge helpful comments from Salvatore Lo Bue, Carl Ceulemans, Lode Lauwaert and the members of the Chair Ethics of AI at KU Leuven.

Author contributions Co-authored, with both authors contributing equally to the manuscript.

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.

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Asaro, P.: On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. Int. Rev. Red Cross. **94**(886), 687–709 (2012). https://doi.org/10.1017/S1816383112000768
2. Dremliuga, R.: General legal limits of the application of the Lethal Autonomous weapons systems within the Purview of International Humanitarian Law. J. Politics Law. **13**(2), 115 (2020). https://doi.org/10.5539/jpl.v13n2p115
3. Egeland, K.: Lethal Autonomous Weapon Systems under International Humanitarian Law. Nordic J. Int. Law. **85**(2), 89–118 (2016). https://doi.org/10.1163/15718107-08502001
4. Grand-Clément, S.: *Artificial Intelligence Beyond Weapons: Application and impact of AI in the military domain*. UNIDIR. (2023). https://unidir.org/publication/artificial-intelligence-beyond-weapons-application-and-impact-of-ai-in-the-military-domain/
5. Van Severen, S., Vander Maelen, C.: Killer robots: Lethal autonomous weapons and international law. In J. de Bruyne & C. Vanleenhove (Eds.), *Artificial intelligence and the law* (pp. 151–172). Intersentia. (2021)
6. Sparrow, R.: Killer Robots. J. Appl. Philos. **24**(1), 62–77 (2007). https://doi.org/10.1111/j.1468-5930.2007.00346.x
7. Crootof, R.: War torts: Accountability for autonomous weapons. Univ. Pa. Law Rev. **164**(6), 1347–1402 (2016)
8. Roff, H.M.: 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). (2014). https://doi.org/10.5860/CHOICE.51-3176
9. Laura, A.D.: Lethal Autonomous weapons systems: The overlooked importance of administrative accountability. In: Alcala, R.T.P., Eric Talbot Jensen (eds.) The Impact of Emerging Technologies on the Law of Armed Conflict, p. 27. Oxford University Press (2019)
10. Vallor, S., Vierkant, T.: Find the gap: AI, responsible Agency and Vulnerability. Minds Machines. **34**(20) (2024). https://doi.org/10.1007/s11023-024-09674-0
11. Christie, E.H., Ertan, A., Adomaitis, L., Klaus, M.: Regulating lethal autonomous weapon systems: Exploring the challenges of explainability and traceability. AI Ethics. (2023). https://doi.org/10.1007/s43681-023-00261-0
12. Lagioia, F., Sartor, G.: AI systems under Criminal Law: A legal analysis and a Regulatory Perspective. Philos. Technol. **33**(3), 433–465 (2020). https://doi.org/10.1007/s13347-019-00362-x
13. List, C.: Group Agency and Artificial Intelligence. Philos. Technol. **34**(4), 1213–1242 (2021). https://doi.org/10.1007/s13347-021-00454-7
14. Simmler, M., Markwalder, N.: Guilty Robots?– rethinking the nature of culpability and legal personhood in an age of Artificial Intelligence. Crim. Law Forum. **30**(1), 1–31 (2019). https://doi.org/10.1007/s10609-018-9360-0
15. Tigard, D.W.: Artificial Moral responsibility: How we can and cannot hold machines responsible. Camb. Q. Healthc. Ethics. **30**(3), 435–447 (2021). https://doi.org/10.1017/S0963180120000985
16. Himmelreich, J., Köhler, S.: Responsible AI through conceptual Engineering. Philos. Technol. **35**(3), 60 (2022). https://doi.org/10.1007/s13347-022-00542-2
17. Champagne, M., Tonkens, R.: Bridging the responsibility gap in Automated Warfare. Philos. Technol. **28**(1), 125–137 (2015). https://doi.org/10.1007/s13347-013-0138-3
18. Boutin, B.: Legal Questions Related to the Use of Autonomous Weapon Systems. Asser Institute: (2021). https://www.asser.nl/media/795707/boutin-legal-questions-related-to-the-use-of-aws.pdf
19. Chengeta, T.: Accountability gap: Autonomous weapon systems and modes of responsibility in international law. Denver J. Int. Law Policy, **45**(1). (2016)
20. Saxon, D.: 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. (2016). https://doi.org/10.4324/9781315578187
21. Mettraux, G.: The evolution of the Law of Command responsibility and the Principle of Legality. In: The Law of Command Responsibility. Oxford University Press (2009b)
22. Himmelreich, J.: Responsibility for Killer Robots. Ethical Theory Moral. Pract. **22**(3), 731–747 (2019). https://doi.org/10.1007/s10677-019-10007-9
23. Nyholm, S.: Attributing Agency to Automated systems: Reflections on human–Robot collaborations and responsibility-loci. Sci Eng. Ethics. **24**(4), 1201–1219 (2018). https://doi.org/10.1007/s11948-017-9943-x
24. Schulzke, M.: Autonomous weapons and distributed responsibility. Philos. Technol. **26**(2), 203–219 (2013). https://doi.org/10.1007/s13347-012-0089-0
25. 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
26. Acquaviva, G.: Autonomous weapons systems controlled by Artificial Intelligence: A conceptual Roadmap for International

Criminal responsibility. SSRN Electron. J. (2021). https://doi.org/10.2139/ssrn.4070447
27. Spadaro, A.: A Weapon is no subordinate. J. Int. Criminal Justice. mqad025 (2023). https://doi.org/10.1093/jicj/mqad025
28. Jessberger, F., Werle, G.: *Principles of international criminal law* (Fourth edition). Oxford University Press. (2020)
29. Kühne, R., Peter, J.: Anthropomorphism in human–robot interactions: A multidimensional conceptualization. Communication Theory. **33**(1), 42–52 (2023). https://doi.org/10.1093/ct/qtac020
30. Garreau, J.: Bots on the Ground. *Washington Post*. (2007)., May 6 https://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009.html
31. Singer, P.W.: Wired for war: The Robotics Revolution and Conflict in the twenty-first Century. Penguin Books (2010)
32. Roesler, E., Manzey, D., Onnasch, L.: A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. Sci. Rob. **6**(58), eabj5425 (2021). https://doi.org/10.1126/scirobotics.abj5425
33. Nijssen, S.R.R., Müller, B.C.N., Baaren, R.B.V., Paulus, M.: Saving the Robot or the human? Robots who feel deserve Moral Care. Soc. Cogn. **37**(1), 41–S2 (2019). https://doi.org/10.1521/soco.2019.37.1.41
34. Kahn, P.H., Kanda, T., Ishiguro, H., Gill, B.T., Ruckert, J.H., Shen, S., Gary, H.E., Reichert, A.L., Freier, N.G., Severson, R.L.: Do people hold a humanoid robot morally accountable for the harm it causes? *Proceedings of the Seventh* Annual *ACM/IEEE International Conference on Human-Robot Interaction*, 33–40. (2012). https://doi.org/10.1145/2157689.2157696
35. Kim, T., Hinds, P.: Who should I blame? Effects of Autonomy and transparency on attributions in Human-Robot Interaction. ROMAN 2006 - 15th IEEE Int. Symp. Robot Hum. Interact. Communication. 80–85 (2006). https://doi.org/10.1109/ROMAN.2006.314398
36. Kneer, M.: Can a Robot Lie? Exploring the Folk Concept of lying as Applied to Artificial agents. Cogn. Sci. **45**(10), e13032 (2021). https://doi.org/10.1111/cogs.13032
37. Kneer, M., Stuart, M.T.: Playing the blame game with Robots. Companion 2021 ACM/IEEE Int. Conf. Human-Robot Interact. 407–411 (2021). https://doi.org/10.1145/3434074.3447202
38. Kneer, M., Christen, M.: Responsibility gaps and retributive dispositions: Evidence from the US, Japan and Germany. SSRN Electron. J. (2023). https://doi.org/10.2139/ssrn.4394118
39. Lima, G., Grgić-Hlača, N., Cha, M.: Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17. (2021). https://doi.org/10.1145/3411764.3445260
40. Liu, P., Du, Y.: Blame attribution asymmetry in human–automation Cooperation. Risk Anal. **42**(8), 1769–1783 (2022). https://doi.org/10.1111/risa.13674
41. Malle, B.F., Magar, S.T., Scheutz, M.: AI in the Sky: How people morally evaluate Human and Machine decisions in a Lethal Strike Dilemma. In: Aldinhas Ferreira, M.I., Silva Sequeira, J., Singh Virk, G., Tokhi, M.O., Kadar, E.E. (eds.) Robotics and Well-Being, vol. 95, pp. 111–133. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-12524-0_11
42. Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J., Cusimano, C.: Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 117–124. (2015). https://doi.org/10.1145/2696454.2696458
43. Malle, B.F., Scheutz, M., Forlizzi, J., Voiklis, J.: Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. 2016 11th ACM/IEEE Int. Conf. Human-Robot Interact. (HRI). **125–132** (2016). https://doi.org/10.1109/HRI.2016.7451743
44. Stuart, M.T., Kneer, M.: Guilty Artificial minds: Folk attributions of Mens Rea and Culpability to Artificially Intelligent agents. Proc. ACM Hum Comput Interact. **5**(CSCW2), 1–27 (2021). https://doi.org/10.1145/3479507
45. Van Der Woerdt, S., Haselager, P.: When robots appear to have a mind: The human perception of machine agency and responsibility. New Ideas Psychol. **54**, 93–100 (2019). https://doi.org/10.1016/j.newideapsych.2017.11.001
46. Voiklis, J., Kim, B., Cusimano, C., Malle, B.F.: Moral judgments of human vs. Robot agents. *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 775–780. (2016). https://doi.org/10.1109/ROMAN.2016.7745207
47. Furlough, C., Stokes, T., Gillan, D.J.: Attributing blame to Robots: I. The influence of Robot Autonomy. Hum. Factors: J. Hum. Factors Ergon. Soc. **63**(4), 592–602 (2021). https://doi.org/10.1177/0018720819880641
48. Caspar, E.A., Ioumpa, K., Arnaldo, I., Di Angelis, L., Gazzola, V., Keysers, C.: Commanding or being a simple intermediary: How does it affect moral behavior and related brain mechanisms? [Preprint] Neurosci. (2021). https://doi.org/10.1101/2021.12.10.472075
49. Moretto, G., Walsh, E., Haggard, P.: Experience of agency and sense of responsibility. Conscious. Cogn. **20**(4), 1847–1854 (2011). https://doi.org/10.1016/j.concog.2011.08.014
50. Salatino, A., Prevel, A., Caspar, E., Lo Bue, S.: Fire! Do not fire! A new paradigm testing how autonomous systems affect agency and moral decision-making. Pre-print. bioRxiv. 2023–2012 (2023). https://doi.org/10.1101/2023.12.19.572326
51. Burin, D., Pyasik, M., Salatino, A., Pia, L.: That's my hand! Therefore, that's my willed action: How body ownership acts upon conscious awareness of willed actions. Cognition. **166**, 164–173 (2017). https://doi.org/10.1016/j.cognition.2017.05.035
52. Haggard, P.: Sense of agency in the human brain. Nat. Rev. Neurosci. **18**(4), 196–207 (2017). https://doi.org/10.1038/nrn.2017.14
53. Jeannerod, M.: The mechanism of self-recognition in humans. Behav. Brain. Res. **142**(1–2), 1–15 (2003). https://doi.org/10.1016/S0166-4328(02)00384-4
54. Pyasik, M., Salatino, A., Burin, D., Berti, A., Ricci, R., Pia, L.: Shared neurocognitive mechanisms of attenuating self-touch and illusory self-touch. Soc. Cognit. Affect. Neurosci. **14**(2), 119–127 (2019). https://doi.org/10.1093/scan/nsz002
55. Haggard, P., Tsakiris, M.: The experience of agency: Feelings, judgments, and responsibility. Curr. Dir. Psychol. Sci. **18**(4), 242–246 (2009). https://doi.org/10.1111/j.1467-8721.2009.01644.x
56. Caspar, E.A., Christensen, J.F., Cleeremans, A., Haggard, P.: Coercion changes the sense of agency in the human brain. Curr. Biol. **26**(5), 585–592 (2016). https://doi.org/10.1016/j.cub.2015.12.067
57. Bandura, A.: Toward a psychology of human agency. Perspect. Psychol. Sci. **1**, 164–180 (2006)
58. Haggard, P., Clark, S., Kalogeras, J.: Voluntary action and conscious awareness. Nat. Neurosci. **5**(4), 382–385 (2002). https://doi.org/10.1038/nn827
59. Moore, J.W., Obhi, S.S.: Intentional binding and the sense of agency: A review. Conscious. Cogn. **21**(1), 546–561 (2012). https://doi.org/10.1016/j.concog.2011.12.002
60. Christensen, J.F., Di Costa, S., Beck, B., Haggard, P.: I just lost it! Fear and anger reduce the sense of agency: A study using intentional binding. Exp. Brain Res. **237**, 1205–1212 (2019). https://doi.org/10.1007/s00221-018-5461-6
61. Imaizumi, S., Tanno, Y.: Intentional binding coincides with explicit sense of agency. Conscious. Cogn. **67**, 1–15 (2019). https://doi.org/10.1016/j.concog.2018.11.005

62. Malik, R.A., Obhi, S.S.: Social exclusion reduces the sense of agency: Evidence from intentional binding. Conscious. Cogn. **71**, 30–38 (2019). https://doi.org/10.1016/j.concog.2019.03.004
63. Blackwood, N.J., Bentall, R.P., Simmons, A., Murray, R.M., Howard, R.J.: Self-responsibility and the self-serving bias: An fMRI investigation of causal attributions. NeuroImage. **20**(2), 1076–1085 (2003). https://doi.org/10.1016/S1053-8119(03)00331-8
64. Wegner, D.M., Wheatley, T.: Apparent mental causation: Sources of the experience of will. Am. Psychol. **54**(7), 480 (1999). https://doi.org/10.1037/0003-066X.54.7.480
65. Caspar, E.A., Cleeremans, A., Haggard, P.: Only giving orders? An experimental study of the sense of agency when giving or receiving commands. PLOS ONE. **13**(9), e0204027 (2018). https://doi.org/10.1371/journal.pone.0204027
66. Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., Gray, K.: Algorithmic discrimination causes less moral outrage than human discrimination. J. Exp. Psychol. Gen. **152**(1), 4–27 (2023). https://doi.org/10.1037/xge0001250
67. Shank, D. B., DeSanti, A., Maninger, T.: When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. Inf. Commun. Soc. **22**(5), 648–663 (2019). https://doi.org/10.1080/1369118X.2019.1568515
68. de Jong, R.: The retribution-gap and responsibility-loci related to Robots and Automated technologies: A reply to Nyholm. Sci Eng. Ethics. **26**(2), 727–735 (2020). https://doi.org/10.1007/s11948-019-00120-4
69. Matthias, A.: The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics Inf. Technol. **6**(3), 175–183 (2004). https://doi.org/10.1007/s10676-004-3422-1
70. de Santoni, F., Mecacci, G.: Four responsibility gaps with Artificial Intelligence: Why they Matter and how to address them. Philos. Technol. (2021). https://doi.org/10.1007/s13347-021-00450-x
71. Hindriks, F., Veluwenkamp, H.: The risks of autonomous machines: From responsibility gaps to control gaps. Synthese. **201**(1), 21 (2023). https://doi.org/10.1007/s11229-022-04001-5
72. Hume, D.: *A treatise of human nature*. Clarendon Press. (1739). https://oll.libertyfund.org/title/bigge-a-treatise-of-human-nature
73. Feier, T., Gogoll, J., Uhl, M.: Hiding behind machines: Artificial agents May help to evade punishment. Sci Eng. Ethics. **28**(2), 19 (2022). https://doi.org/10.1007/s11948-022-00372-7
74. Greene, J.: From neural 'is' to moral 'ought': What are the moral implications of neuroscientific moral psychology? Nat. Rev. Neurosci. **4**(10), 846–850 (2003)