Autonomous Weapons and Ethical Judgments

Experimental Evidence on Attitudes toward the Military Use of "Killer Robots"

Pre-print accepted for publication in *Peace and Conflict: Journal of Peace Psychology*

Ondřej Rosendorf
Charles University
Peace Research Center Prague
Experimental Lab for
International Security Studies
Prague, Czech Republic
ondrej.rosendorf@fsv.cuni.cz

Michal Smetana
Charles University
Peace Research Center Prague
Experimental Lab for
International Security Studies
Prague, Czech Republic
michal.smetana@fsv.cuni.cz

Marek Vranka
Charles University
Peace Research Center Prague
Experimental Lab for
International Security Studies
Prague, Czech Republic
marek.vranka@fsv.cuni.cz

Abstract

The advent of autonomous weapons brings intriguing opportunities and significant ethical dilemmas. This article examines how increasing weapon autonomy affects approval of military strikes resulting in collateral damage, perception of their ethicality, and blame attribution for civilian fatalities. In our experimental survey of U.S. citizens, we presented participants with scenarios describing a military strike with the employment of weapon systems with different degrees of autonomy. The results show that as weapon autonomy increases, the approval and perception of the ethicality of a military strike decreases. However, the level of blame towards commanders and operators involved in the strike remains constant regardless of the degree of autonomy. Our findings suggest that public attitudes to military strikes are, to an extent, dependent on the level of weapon autonomy. Yet, in the eyes of ordinary citizens, this does not take away the moral responsibility for collateral damage from human entities as the ultimate "moral agents".

Keywords

autonomous weapons; killer robots; survey experiment; public attitudes

Public Significance Statement

This study examines differences in public perceptions of autonomous weapons – one of the key military innovations of our time. We demonstrate that the public perceives the use of fully autonomous weapon systems as more ethically problematic than systems with lower autonomy.

Introduction

The advent of autonomous weapon systems (AWS), also known as “killer robots”,[1] brings intriguing opportunities and significant ethical dilemmas. In simple terms, AWS are weapons that allow selecting and engaging targets without human intervention (ICRC, 2016). In some limited capacity, these weapons already exist. For example, some loitering munitions and air defense systems used by Israel and the United States could qualify as AWS (Horowitz, 2016a; Sauer, 2021). Militaries around the world invest in weapon autonomy because it promises a speed-based advantage in combat, reduced reliance on communication links, or decreased labor demands (Horowitz, 2019; ICRC, 2021). However, it also raises serious strategic, legal, and ethical concerns.

Perhaps the most controversial issue in the current debate is the possibility that AWS may be entrusted with decisions to end human life (Bode & Huelss, 2018). Many scholars believe that delegating lethal decision-making to machines is fundamentally unethical because doing so would impinge on the right to life and human dignity of affected persons (Asaro, 2012; Rosert & Sauer, 2021). Another part of the scholarship worries that if AWS were to malfunction or commit war crimes, there would be arguably no one who could be justly held responsible for such outcomes, resulting in “responsibility gaps” (Matthias, 2004; Sparrow, 2007). In this context, AWS present certain unique ethical challenges.

Because of the troubling direction of military-technological development, opponents of AWS formed the “Campaign to Stop Killer Robots”, a coalition of non-governmental organizations advocating for a ban on fully autonomous weapons (CSKR, n.d.). Advocacy efforts have helped bring the issue to the attention of the international community, which has been discussing weapon autonomy since 2013 under the United Nations (UN) Convention on

[1] While the term “killer robots” remains widely popular in the public discourse, some scholars point out that its use gives the subject matter an inappropriate “sci-fi feel”, which feeds into the dark imagination of AWS as “Terminators” (Rosert & Sauer, 2021; Young & Carpenter, 2018).

Certain Conventional Weapons (CCW). In 2017, States Parties to the convention established a Group of Governmental Experts (GGE) on “lethal autonomous weapons systems”, a subsidiary body of the CCW tasked to formulate recommendations on how to address AWS (Bahçecik, 2019; Bolton & Mitchell, 2020). The UN Secretary-General previously urged the GGE “to deliver”, arguing that it would be “morally repugnant” if the world fails to ban such weapons (Bugge, 2018; UN News, 2019). Some 30 countries have already expressed their support for a legally binding instrument (Human Rights Watch, 2020), but the feasibility of such an outcome remains uncertain at this point (Rosendorf, 2021; Rosert & Sauer, 2021).

According to some experts, the prospects of international control, or even prohibition, of AWS depend, to an extent, on the public opposition to these systems and their perceived unethicality (Scharre, 2018; Young & Carpenter, 2018). Some go as far as to argue that the use of AWS despite public opposition would violate the so-called Martens Clause, which prohibits the use of weapons contrary to the “dictates of public conscience” (Human Rights Watch, 2012). From an ethical standpoint, autonomous weapons have certain qualities that might make the public see them as relatively more abhorrent than other (non-autonomous) weapons. There is, however, still a lack of scholarly research that would examine public attitudes to AWS from a comparative perspective.

Most of the existing surveys on the topic are merely descriptive (Carpenter, 2013; Galliott & Wyatt, 2020; Ipsos, 2019; Moshkina & Arkin, 2008; Van der Loos & Croft, 2015). The few survey experiments examine issues such as the impact of popular culture on public perceptions (Young & Carpenter, 2018), the varying levels of public support in the context of increasing military utility and development patterns in foreign countries (Horowitz, 2016b), the relationship between autonomy and accountability for civilian casualties (Walsh, 2015), or the acceptability of military applications of artificial intelligence (AI) among AI researchers (Zhang et al., 2021). However, there is currently no experimental research that would examine the relationship between varying degrees of weapon autonomy and ethical judgments.

To fill this gap, we conducted a survey experiment on a sample of 1,006 U.S. citizens. The survey examined how different degrees of weapon autonomy affect public approval of military strikes resulting in collateral damage, perception of their ethicality, and attribution of blame for civilian fatalities. First, we presented our participants with fictional scenarios describing military strikes involving weapons with varying degrees of autonomy. Subsequently, we asked questions related to the approval and ethicality of these strikes and the moral responsibility of relevant entities for collateral damage.

The results show that as weapon autonomy increases, public approval and perception of the ethicality of military strikes with collateral damage decrease. We also found that increasing weapon autonomy is associated with higher blaming for civilian fatalities, albeit only toward the programmer, manufacturer, and the machine. The amount of blame remained constant in the case of the operator and the commanding officer. Overall, humans rather than machines remain the most blamed entities irrespective of the degree of autonomy.

The remainder of this article proceeds as follows. First, we discuss the ethical dimension of weapon autonomy and formulate our hypotheses. In the next section, we lay out the research design of our survey experiment. We then present the results and follow with a brief discussion of our main findings. Finally, we conclude by suggesting some avenues for further research.

Ethical Dimension of Weapon Autonomy

In a military context, the extent to which AWS present an ethical issue depends mainly on the “degree” of their autonomy and the types of functions being made autonomous. One approach to assessing weapon autonomy considers the degree of human involvement in tasks carried out by the machine. “Human-in-the-loop” systems require human input at some point of task execution. “Human-on-the-loop” systems perform some tasks independently, but their operation is monitored by a human who can intervene. And, finally, “human-out-of-the-loop” systems perform some tasks independently without any human input (Boulanin & Verbruggen, 2017; Scharre & Horowitz, 2015).

A complementary approach looks at the types of functions being made autonomous. While autonomy in functions such as navigation, landing, or refueling is generally accepted as ethically unproblematic, autonomy in target selection and engagement raises serious ethical concerns (Boulanin & Verbruggen, 2017). With most of today’s weapon systems, including remote-controlled armed drones, humans still decide whether to select and engage specific targets (Horowitz, 2017; Scharre, 2018). Nevertheless, there are some notable exceptions that, arguably, already constitute examples of AWS in use.

The Israeli Harpy is a loitering munition that, once launched, detects and attacks enemy radar signatures without human supervision. When the Harpy finds a target that meets the preprogrammed parameters, persons responsible for its launch are “out of the loop” and unable to intervene. Various air defense systems, including the U.S. Phalanx, are also capable of autonomous target selection and engagement. In these cases, however, persons responsible for the system’s operation usually remain “on the loop” to override its programming if the need arises (Horowitz, 2016a; Sauer, 2021). Although the military use of these systems is currently limited, autonomy in target selection- and engagement-related functions has been on the rise (Roff, 2016), and many systems are “only a software update away” from AWS (Bode & Huelss, 2018, p. 400). Concerns about weapon autonomy are, thus, relevant for the present rather than some distant future.

Some of the ethical objections to AWS are equally applicable to remotely controlled systems or long-range weapons. For example, the use of AWS would likely contribute to a greater physical and emotional distancing from the battlefield and thereby erode the “natural compulsion of men not to kill" (Grossman, 1995; Sharkey, 2012, p. 112). Other objections are potentially unique to AWS. One argument emphasizes that ceding lethal decision-making to machines is problematic because no technology can (currently) substitute human judgment, which is necessary for evaluating whether the attack would comply with the provisions of the International Humanitarian Law (IHL) (Asaro, 2012; Sharkey, 2012). However, many scholars argue that machines should not have life-and-death powers in the first place, even if they were technically capable of making IHL judgments. From this standpoint, the shift to unsupervised delivery of lethal force impinges on the right to life and dignity of affected persons by making targeting decisions arbitrary and by reducing human beings to sensor data (Asaro, 2012; ICRC, 2021; Rosert & Sauer, 2019).

It is conceivable that the public shares similar concerns about the ethicality of AWS. The results from previous surveys indicate that this might be the case. In a recent public opinion survey by Ipsos (2019), 61 percent of participants from 26 countries opposed the use of AWS. Of those who opposed, roughly two-thirds agreed that allowing machines to kill would cross a moral line. An earlier survey of robotics researchers found that autonomous robots were the least acceptable entities in warfare compared to soldiers and robots as an extension of a human soldier (Moshkina & Arkin, 2008). We, therefore, expect that the public will be more opposed to the use of weapons with higher degrees of autonomy. Following previous studies on attitudes to military strikes (Press et al., 2013), we use approval and perception of ethicality as our main measures. It is possible that some participants may approve of the strike despite seeing it as unethical, for example, for strategic rather than normative reasons.

Hypothesis 1: With the increasing autonomy of a weapon system, public approval of the military strike and the perception of its ethicality will decrease.

Another objection, which is potentially unique to AWS, concerns issues related to the attribution of moral and legal responsibility for negative outcomes resulting from their use. Some scholars worry that if AWS were to malfunction or commit war crimes, there would be arguably no one who could be justly held responsible, especially if they could not control or foresee how the system would behave. At the same time, the system itself would lack moral agency as a prerequisite for the attribution of responsibility (Asaro, 2012). The use of weapons that select and engage targets autonomously could therefore create "responsibility gaps" or "accountability gaps" (Matthias, 2004; Sparrow, 2007). Consequently, such "gaps" could be exploited by the political and military leadership to escape criminal liability (Human Rights Watch, 2015). Insisting on holding someone responsible would, nevertheless, entail the risk of scapegoating (Liu, 2016) despite that negative outcomes may result from genuine accidents (Dunlap, 2016; Robillard, 2018).

We can identify several entities that could be held responsible, irrespective of the difficulties associated with responsibility attribution. Some of the most frequently mentioned entities in the literature include the commander, operator, programmer, manufacturer, and the machine itself (Cass, 2015; Crootof, 2015; Sparrow, 2007). To be sure, this list is by no means exhaustive. Other potential culprits could include political decision-makers, employers who fund the research and development of the technology, or algorithms, computers, and sensors (Walsh, 2015). For our research purposes, we opt for a limited number of the most frequently mentioned entities.

Some scholars believe that machines could eventually become "artificial moral agents" capable of assuming responsibility for their actions (Wallach & Allen, 2013). Others argue that current-day robots cannot be held responsible because they cannot be punished and suffer as a result (Sharkey, 2012; Sparrow, 2007). However, people tend to assign responsibility even to inanimate entities such as companies, which can be sued and punished. Hellström (2013) observes that such tendency increases with the degree of an entity's autonomy. Some limited evidence from previous surveys also shows that people see machines as blamable (Kim & Hinds, 2006). We, thus, expect that the public will blame the machine more as the autonomy in its target selection and engagement functions increases.

Hypothesis 2a: As the autonomy of a weapon system increases, public blame of the machine for the collateral damage will also increase.

Nevertheless, even if the use of AWS would result in negative outcomes that were unforeseeable, responsibility for the launch alone could be attributed within the chain of command (Dunlap, 2016; Kalmanovitz, 2017). As Schulzke (2013) points out, modern militaries already operate through "distributed responsibility", where commanders share responsibility for the actions of their subordinates. As such, we expect that the public will continue to blame human agents despite the increasing autonomy.

Hypothesis 2b: Notwithstanding the increasing autonomy of a weapon system, human agents (commander, operator, and programmer) will still be blamed more than the machine for the collateral damage.

Research Design

We conducted our survey experiment on a sample of 1,006 U.S. adults on the Amazon Mechanical Turk (AMT) platform, aiming for the power of 0.8 to detect an effect size d = 0.3 when comparing pairs of treatments. This is not a representative sample. The composition of U.S. respondents recruited through AMT is biased toward males, younger, and more liberal and educated respondents (Huff & Tingley, 2015). However, the existing meta-studies show that the results of experiments conducted on the AMT platform are comparable to those relying on representative samples (Clifford et al. 2015). For our research purposes, it is justifiable to use a convenience sample because we are interested in examining the relationship between degrees of autonomy and judgments of ethicality rather than describing the attitudes of the U.S. population writ large (Coppock & McClellan, 2019).[2]

In the survey, we described a fictional military strike on a terrorist hideout near the Pakistani border. We randomly assigned each participant to one of the experimental treatments: the strike was conducted by a helicopter, a remotely controlled drone, a human-supervised autonomous drone, or a fully autonomous drone. In each version of the scenario, we described the degree of autonomy in target selection and engagement. In the first two treatments, the pilot or operator selected and engaged the target ("human-in-the-loop"). In the third version, the drone selected and engaged the target under human supervision ("human-on-the-loop").[3] In the fourth, the drone selected and engaged the target without human involvement ("human-out-of-the-loop"). In each scenario, we also provided information that five civilian bystanders were killed as collateral damage.

On the following page, participants answered how much they approved of the strike (Likert scale from 1 – strongly disapprove to 6 – strongly approve) and how ethical they found it (1 – highly unethical to 6 – highly ethical). They also completed an attention check by selecting the location of the strike (only those participants who answered correctly were allowed to continue). In the next section, participants assigned blame for civilian fatalities (1 – no blame at all to 6 – maximum blame) to each of the following entities: the person who piloted the helicopter or drone, the programmer of the targeting mechanism, the company that

[2] See Supplementary Materials for a more detailed description of the demographic composition of our sample as well as the description of experimental conditions.

[3] There were two additional sub-scenarios based on the "human-on-the-loop" scenario, which were related to a hypothesis about the compliance with the notion of the "meaningful human control". Due to space constraints, these results are not discussed in the main text of this paper, but they can be found in the Supplementary Materials.

manufactured the machine, the machine itself, and the commanding officer. Participants then answered the second attention check item, socio-demographic questions, and a personality questionnaire. Finally, we presented participants with all three remaining versions of the scenario and asked them to rate their approval of the strike.

Results

First, we analyzed the blame for each entity in different scenarios (see Figure 1).[4] The commanding officer was the most blamed entity, and the level of blame did not differ across the scenarios ($F_{(3,\ 653)} = 0.573$, $p = .633$). The operator was the second most blamed entity, and the level of blame was also consistent across the scenarios ($F_{(3,\ 653)} = 1.83$, $p = .141$). In contrast, the blame of the programmer ($F_{(3,\ 653)} = 3.67$, $p = .012$), the company ($F_{(3,\ 653)} = 7.53$, $p < .001$), and the machine ($F_{(3,\ 653)} = 14.0$, $p < .001$) differed significantly between the scenarios. Specifically, the blame for these three entities grew with the increasing level of weapon autonomy (all p-values for linear contrasts < .003).

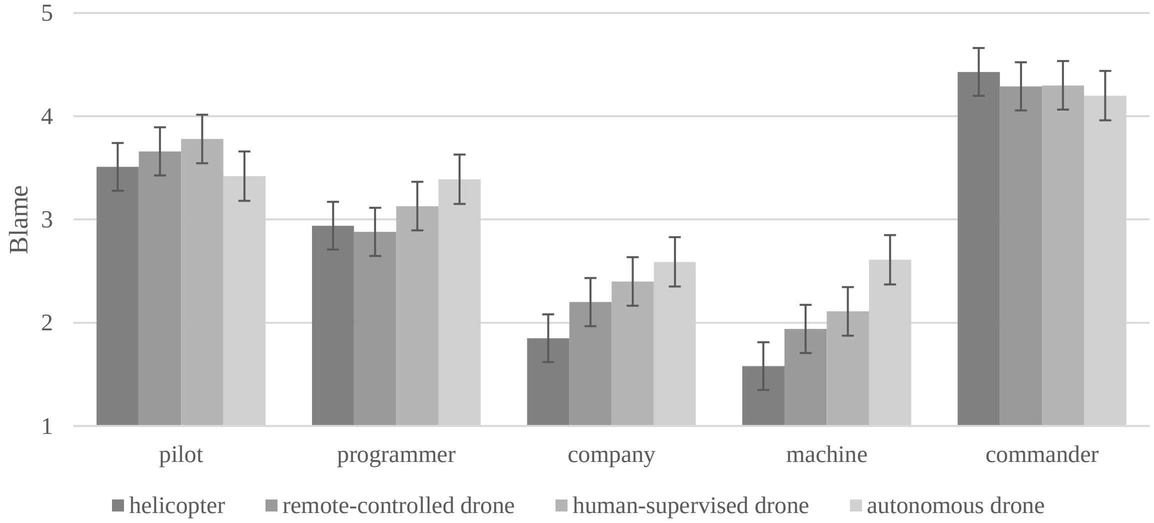

Figure 1: Amount of blame for the collateral damage

[4] See Supplementary Materials for more detailed analyses and results.

Note. All blame judgments were made on 6-point scales (1 – no blame to 6 – maximum blame). Error bars represent 95% CI. *N* = 657.

The results hold even when controlling for education, income, political identification, attitude toward the military, age, and gender. Age was negatively related to blaming of all entities across scenarios. A positive attitude toward the military was significantly related to lesser blaming of all entities, except for the machine. The only other variable significantly related to the amount of blame was political identification: conservatism was related to higher blame of the machine and lower blame of the commanding officer, while liberalism was related to lower blame of the machine and higher blame of the commanding officer.[5]

Second, we used ANCOVAs to analyze whether participants differed in their judgment of ethicality and approval of the strike across the same four scenarios, controlling for the level of education, income, political identification, attitudes toward military, age, and gender. The approval correlated highly with the judgment of ethicality (r = .83), and more conservative participants and those with positive attitudes toward the military judged the strike as more ethical and approved of it more across all scenarios. Conversely, more liberal participants and those with negative attitudes toward the military judged the strike as less ethical and approved of it less across all scenarios. In addition, the strikes in the scenarios with increasing autonomy were judged as less ethical (t_{647} = 2.57, p = .010), and participants approved of them slightly, albeit not significantly, less as well (t_{647} = 1.89, p = .059).

Since participants rated their approval of the strike in all versions of the scenario at the end of the survey, we were also able to analyze the differences in approval between scenarios using a repeated-measures ANOVA (see Figure 2). The results of this test provide further evidence of decreasing approval of scenarios with increasing weapon autonomy ($F_{(2.62, 2636.30)}$

[5] Political identification is a scale ranging from 1 to 6, from very liberal to very conservative.

= 77.5, *p* < .001). We also found that across all four scenarios, participants who perceived strikes as less ethical were more likely to blame the commander more (*r* = -.34), and those who judged the strike as more ethical blamed the machine more (*r* = .17).

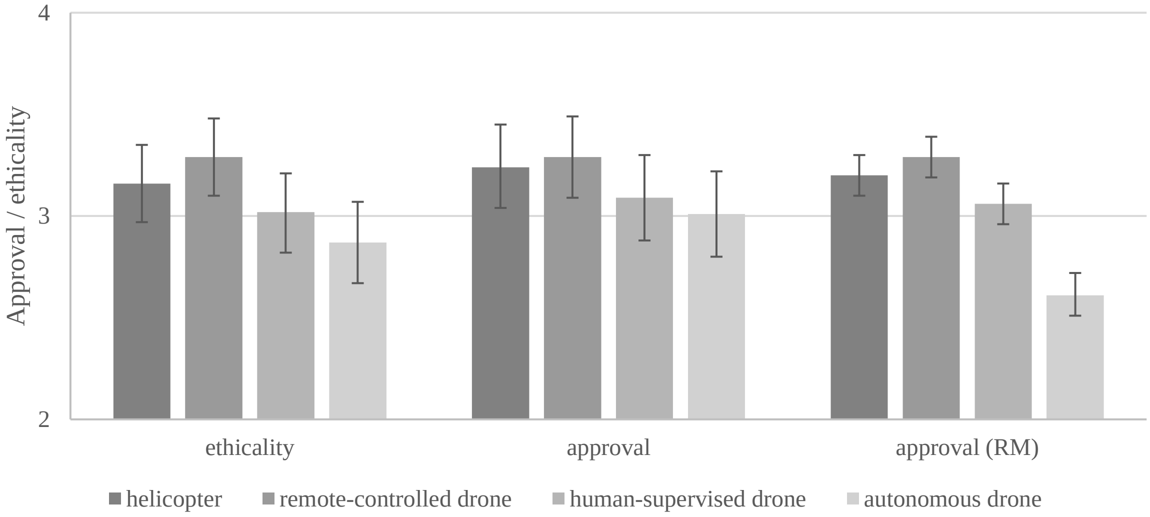

Figure 2: Judgment of ethicality and approval of the strike

Note. The "ethicality" and "approval" judgments are based on evaluations of a single scenario at the beginning of the survey. The "approval (RM)" judgments are based on parallel evaluations of all scenarios made at the end of the survey. All judgments were made on 6-point scales (1 – strongly disapprove / highly unethical to 6 – strongly approve / highly ethical). Error bars represent 95% CI. *N* = 657.

Discussion

The results of our research reveal that increasing autonomy in target selection and engagement functions of weapon systems is associated with lower public approval rates and lower perception of the ethicality of military strikes resulting in collateral damage. The use of the "human-out-of-the-loop" systems, specifically, was the least approved and seen as the least ethical in our survey. The correlation between perception of ethicality and approval also suggests that the public likely disapproves of AWS based on ethical reasoning. While earlier surveys have already indicated the existence of public aversion to AWS (Carpenter, 2013; Ipsos, 2019), our findings link the variance in approval and ethicality to varying degrees of autonomy. Additionally, the finding that political conservatives on balance tend to approve of the use of military force more than political liberals is in line with the results from previous survey experiments (Press et al., 2013; Sagan & Valentino, 2017; Smetana & Vranka, 2021).

The results also shed light on the interplay between moral responsibility and degrees of autonomy. We found that the use of increasingly autonomous weapons leads to higher blaming for civilian fatalities, albeit only toward the programmer of the targeting mechanism, the manufacturer, and the machine. Following our expectations, participants blamed the machine significantly more in the "human-out-of-the-loop" scenario. These findings provide further evidence for the claim that people tend to see machines as blamable (Hellström, 2013; Kim & Hinds, 2006; Walsh, 2015). Nevertheless, our survey design does not allow us to determine whether the amount of blame captures "moral" or merely "causal" responsibility (Liu, 2016; Robillard, 2018). Participants could have blamed the machine more simply because they perceived it more closely linked to collateral damage.

We also found that the amount of blame toward the operator and the commanding officer was unrelated to the degree of weapon autonomy. The level of blame for civilian fatalities attributed to these entities remained constant across all scenarios. In addition, our participants considered human agents to be the most blamable entities. Specifically, the commander, the operator, and the programmer received more blame than the machine in all scenarios, including the "human-out-of-the-loop" scenario. These findings problematize claims that the use of fully autonomous weapons would result in "responsibility gaps" (Matthias, 2004; Sparrow, 2007). In the eyes of ordinary citizens, the use of AWS still does not take away the responsibility for collateral damage from human entities as the ultimate moral agents.

Our findings are in line with Walsh (2015), who observes that the use of AWS does not decrease the degree to which leaders are held responsible for negative outcomes. While the author also finds the tendency of participants to blame the machine more as the autonomy increases, we found a more substantive effect of increasing autonomy on blame attribution toward this entity. One possible explanation is that Walsh asked the participants to attribute the responsibility to a weapon's "sensors and computer" rather than to the "machine" as such. It is conceivable that people tend to see the "machine" as a standalone entity worthy of blame. On the other hand, "sensors and computer" might be perceived as mere components of a weapon system.

Conclusion

In this article, we examined how the use of weapon systems with varying degrees of autonomy affects public approval and perception of the ethicality of military strikes resulting in collateral damage. The evidence suggests that increasing weapon autonomy is associated with lower approval rates and lower perception of ethicality. These findings have potentially important implications for the current discussion about the possibility of international control of AWS at the UN. When it comes to military strikes with collateral damage, the public perceives the use of AWS as relatively more unethical than conventional inhabited and remote-controlled systems. The failure by States Parties to the CCW to reflect adequately on ethical challenges posed by weapon autonomy might, therefore, result in public backlash.

We also investigated the relationship between degrees of weapon autonomy and blame for collateral damage. Similarly to Walsh (2015), we found that, in the eyes of the public, the use of increasingly autonomous weapons does not take away the moral responsibility for negative outcomes from human entities. For our participants, AWS do not mark a qualitative shift in blame attribution. One of the potential implications is that, rather than resulting in "responsibility gaps" as described by Matthias (2004) and Sparrow (2007), the use of AWS could create a gap between the moral and legal responsibility if delegating lethal decision-making to machines would make it easier for the involved persons to escape liability.

Some limitations of our research can serve as potential avenues for future studies on the topic. For example, one of the reasons why our participants saw AWS as relatively more unethical could be due to their novelty. It is also plausible that the participants saw human entities as the most blamable because they have limited experience in dealing with autonomous machines. Technological advances in this area will likely affect how the public feels about these issues. Future research could focus on specific factors, such as sensitivity to the loss of agency and human dignity (Asaro, 2012; Rosert & Sauer, 2019), which might help to explain the differences in perceived ethicality across the spectrum of weapon autonomy. Furthermore, we have only examined public attitudes in the context of military strikes resulting in collateral damage. Future research could investigate scenarios without civilian fatalities. Finally, some elite groups, including the military and political decision-makers, might differ in their ethical judgments regarding AWS. This opens another intriguing avenue for future research.

References

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

Bahçecik, Ş. O. (2019). Civil society responds to the AWS: Growing activist networks and shifting frames. *Global Policy*, *10*(3), 365–369. https://doi.org/10.1111/1758-5899.12671

Bode, I., & Huelss, H. (2018). Autonomous weapons systems and changing norms in international relations. *Review of International Studies*, *44*(3), 393–413. https://doi.org/10.1017/S0260210517000614

Bolton, M. B., & Mitchell, C. C. (2020). When scientists become activists: The International Committee for Robot Arms Control and the politics of killer robots. In M. B. Bolton,

S. Njeri & Benhamin-Britton, T. (Eds.), *Global activism and humanitarian disarmament* (pp. 27–58). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-27611-9_2

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

Bugge, A. (2018, November 5). *U.N.'s Guterres urges ban on autonomous weapons.* Reuters. https://www.reuters.com/article/us-portugal-websummit-un/u-n-s-guterres-urges-ban-onautonomous- weapons-idUSKCN1NA2HG

Carpenter, C. (2013, June 19). *How do Americans feel about fully autonomous weapons?* The Duck of Minerva. https://duckofminerva.com/2013/06/how-do-americans-feel-about-fully-autonomous-weapons.html

Cass, K. (2015). Autonomous weapons and accountability: Seeking solutions in the law of war. *Loyola of Los Angeles Law Review*, *48*(3), 1017–1067.

Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology? *Research & Politics*, *2*(4), 1–9. https://doi.org/10.1177/2053168015622072

Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics*, *6*(1), 1–14. https://doi.org/10.1177/2053168018822174

Crootof, R. (2015). The killer robots are here: legal and policy implications. *Cardozo Law Review*, *36*(5), 1837–1915.

CSKR. (n.d.). *The story so far*. Retrieved November 18, 2021, from https://www.stopkillerrobots.org/about/

Dunlap, C. J. (2016). Accountability and autonomous weapons: Much ado about nothing? *Temple International and Comparative Law Journal*, *30*(1), 63–76.

Galliott, J., & Wyatt, A. (2020). Risks and benefits of autonomous weapon systems: Perceptions among future Australian Defence Force officers. *US Air Force Journal of Indo-Pacific Affairs*, *3*(4), 17–34.

Grossman, D. (1995). *On killing: The psychological cost of learning to kill in war and society*. Little, Brown and Co.

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

Horowitz, M. C. (2016a). Why words matter: The real world consequences of defining autonomous weapons systems. *Temple International and Comparative Law Journal*, *30*(1), 85–98.

Horowitz, M. C. (2016b). Public opinion and the politics of the killer robots debate. *Research & Politics*, *3*(1), 1–8. https://doi.org/10.1177%2F2053168015627183

Horowitz, M. C. (2017). Military robotics, autonomous systems, and the future of military effectiveness. In D. Reiter (Ed.), *The sword's other edge: Trade-offs in the pursuit of military effectiveness* (pp. 161–196). Cambridge University Press. https://doi.org/10.1017/9781108241786.006

Horowitz, M. C. (2019). When speed kills: Lethal autonomous weapon systems, deterrence and stability. *Journal of Strategic Studies*, *42*(6), 764–788. https://doi.org/10.1080/01402390.2019.1621174

Huff, C., & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, *2*(3), 1–12. https://doi.org/10.1177/2053168015604648

Human Rights Watch. (2012). *Losing humanity: Case against killer robots*. https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots

Human Rights Watch. (2015). *Mind the gap: Lack of accountability for killer robots*. https://www.hrw.org/sites/default/files/reports/arms0415_ForUpload_0.pdf

Human Rights Watch. (2020, August 10). *Stopping killer robots: Country positions on banning fully autonomous weapons and retaining human control*. https://www.hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and

ICRC. (2016). *Autonomous weapon systems: Implications of increasing autonomy in the critical functions of weapons*. https://shop.icrc.org/autonomous-weapon-systems-implications-of-increasing-autonomy-in-the-critical-functions-of-weapons-print-en

ICRC. (2021). *ICRC position on autonomous weapon systems*. https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems

Ipsos. (2019, January 22). *Six in ten (61%) respondents across 26 countries oppose the use of lethal autonomous weapons systems*. https://www.ipsos.com/en-us/news-polls/human-rights-watch-six-in-ten-oppose-autonomous-weapons

Kalmanovitz, P. (2017). Lethal autonomous weapons systems and the risks of 'riskless warfare'. In R. Geiss (Ed.), *Lethal autonomous weapons systems: Technology, definition, ethics, law and security* (pp. 184–195). German Federal Foreign Office.

Kim, T., & Hinds, P. (2006). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. *ROMAN 2006 – The 15th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 80–85). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ROMAN.2006.314398

Liu, H. (2016). Refining responsibility: differentiating two types of responsibility issues raised by autonomous weapons systems. In N. Bhuta, S. Beck, R. Geiss, H. Liu & C.

Kreiss (Eds.), *Autonomous weapons systems: Law, ethics, policy* (pp. 325–344). Cambridge University Press. https://doi.org/10.1017/CBO9781316597873.014

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

Moshkina, L., & Arkin, R. (2008). *Lethality and Autonomous Systems: Survey Design and Results*. Georgia Institute of Technology. http://hdl.handle.net/1853/20068

Press, D. G., Sagan, S. D., & Valentino, B. A. (2013). Atomic aversion: Experimental evidence on taboos, traditions, and the non-use of nuclear weapons. *American Political Science Review*, *107*(1), 188–206. https://doi.org/10.1017/S0003055412000597

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. (2016, September 28). *Weapons autonomy is rocketing*. Foreign Policy. https://foreignpolicy.com/2016/09/28/weapons-autonomy-is-rocketing/

Rosendorf, O. (2021). Predictors of support for a ban on killer robots: Preventive arms control as an anticipatory response to military innovation. *Contemporary Security Policy*, *42*(1), 30–52. https://doi.org/10.1080/13523260.2020.1845935

Rosert, E., & Sauer, F. (2019). Prohibiting autonomous weapons: Put human dignity first. *Global Policy*, *10*(3), 370–375. https://doi.org/10.1111/1758-5899.12691

Rosert, E., & Sauer, F. (2021). How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies. *Contemporary Security Policy*, *42*(1), 4–29. https://doi.org/10.1080/13523260.2020.1771508

Sagan, S. D., & Valentino, B. A. (2017). Revisiting Hiroshima in Iran: What Americans really think about using nuclear weapons and killing noncombatants. *International Security*, *42*(1), 41–79. https://doi.org/10.1162/ISEC_a_00284

Sauer, F. (2021). Stepping back from the brink: Why multilateral regulation of autonomy in weapons systems is difficult, yet imperative and feasible. *International Review of the Red Cross*, *102*(913), 235–259. https://doi.org/10.1017/S1816383120000466

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. W. W. Norton & Company.

Scharre, P., & Horowitz, M. C. (2015). *An introduction to autonomy in weapon systems*. Center for a New American Security. https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-weapon-systems

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

Sharkey, N. E. (2012). Killing made easy: From joysticks to politics. In P. Lin, K. Abney & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 111–128). MIT Press.

Smetana, M., & Vranka, M. (2021). How moral foundations shape public approval of nuclear, chemical, and conventional strikes: New evidence from experimental surveys. *International Interactions*, *47*(2), 374–390. https://doi.org/10.1080/03050629.2020.1848825

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

UN News. (2019, March 25). *Autonomous weapons that kill must be banned, insists UN chief.* https://news.un.org/en/story/2019/03/1035381

Van der Loos, M. H. F., & Croft, E. (2015). *The ethics and governance of lethal autonomous weapons systems: An international public opinion poll*. Open Roboethics initiative. http://www.openroboethics.org/wp-content/uploads/2015/11/ORi_LAWS2015.pdf

Wallach, W., & Allen, C. (2013). Framing robot arms control. *Ethics of Information Technology*, *15* (2), 125–135. https://doi.org/10.1007/s10676-012-9303-0

Walsh, J. I. (2015). Political accountability and autonomous weapons. *Research & Politics*, *2*(4), 1–6. https://doi.org/10.1177%2F2053168015606749

Young, K. L., & Carpenter, C. (2018). Does science fiction affect political fact? Yes and no: A survey experiment on "killer robots". *International Studies Quarterly*, *62*(3), 562–576. https://doi.org/10.1093/isq/sqy028

Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2021). Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. *Journal of Artificial Intelligence Research*, *71*(2021), 591–666. https://doi.org/10.1613/jair.1.12895

Supplementary Materials

Data collection

We continued the data collection until the number of participants exceeded 1,000, after the exclusion of those participants who failed to correctly answer all attention-check items. Our final sample has the following socio-demographic characteristics: 52.7% male, median age 34 years, median household income before taxes between $50,000 and $60,000, 63.6% having a bachelor's or higher university degree, 44.3% identifying as Liberals, 22.5% as Moderates, and 33.2% as Conservatives, 77.2% having slightly, moderately, or strongly positive feelings toward the U.S. military.[6] In addition to the experimental groups described in the main text we presented participants with two additional sub-scenarios, Scenario 4 and 5, described below.

Full description of scenarios

Note. We use armed drones as a reference subject for the different degrees of autonomy for two main reasons. First, we expect participants to be more familiar with armed drones than weapons such as loitering munitions, robotic sentry guns, and so on. Second, we want to present participants with plausible military uses of autonomy, rather than engaging fantasies about Terminators, and other misleading depictions of "killer robots".

Scenario 1 – Helicopter

Last week, the military carried out a strike on a terrorist hideout in a village located in the mountain area on the Pakistani border. The strike aimed at the hideout of the local terrorist group was conducted by a helicopter, which was operated by a pilot who selected and fired at

[6] For a discussion of the Amazon Mechanical Turk platform as a research survey tool, see Dupuis et al. (2013).

the target. While neutralizing the terrorist threat, the strike also caused collateral damage, killing 5 civilian bystanders.

Scenario 2 – Remotely-controlled drone

Last week, the military carried out a strike on a terrorist hideout in a village located in the mountain area on the Pakistani border. The strike aimed at the hideout of the local terrorist group was conducted by a drone, which was remotely controlled by an operator who selected and fired at the target. While neutralizing the terrorist threat, the strike also caused collateral damage, killing 5 civilian bystanders.

Scenario 3 – Human-supervised autonomous drone

Last week, the military carried out a strike on a terrorist hideout in a village located in the mountain area on the Pakistani border. The strike aimed at the hideout of the local terrorist group was conducted by a drone, which autonomously selected and fired at the target while an operator monitored its actions. While neutralizing the terrorist threat, the strike also caused collateral damage, killing 5 civilian bystanders.

Scenario 4 – Human-supervised autonomous drone, MHC upheld

Last week, the military carried out a strike on a terrorist hideout in a village located in the mountain area on the Pakistani border. The strike aimed at the hideout of the local terrorist group was conducted by a drone, which autonomously selected and fired at the target while an operator monitored its actions. The operator was technically capable of stopping the drone from firing but made a conscious decision not to intervene after timely and informed consideration of the situation. While neutralizing the terrorist threat, the strike also caused collateral damage, killing 5 civilian bystanders.

Scenario 5 – Human-supervised autonomous drone, MHC violated

Last week, the military carried out a strike on a terrorist hideout in a village located in the mountain area on the Pakistani border. The strike aimed at the hideout of the local terrorist group was conducted by a drone, which autonomously selected and fired at the target while an operator monitored its actions. The operator was technically capable of stopping the drone from firing but did not have the time or information to consider the situation consciously and decide whether to intervene. While neutralizing the terrorist threat, the strike also caused collateral damage, killing 5 civilian bystanders.

Scenario 6 – Fully autonomous drone

Last week, the military carried out a strike on a terrorist hideout in a village located in the mountain area on the Pakistani border. The strike aimed at the hideout of the local terrorist group was conducted by a drone, which autonomously selected and fired at the target without any human input so that an operator was technically incapable of aborting the attack. While neutralizing the terrorist threat, the strike also caused collateral damage, killing 5 civilian bystanders.

Analysis 1 – amount of blame for the collateral damage ascribed to different entities

We performed a 2-factor mixed ANOVA with the evaluated entity as the within-subject factor and the scenario as the between-subject factor. As the test of sphericity was significant (Mauchly's W = .671, p < .001), we used the Greenhouse-Geisser correction (ε = .818) for within-subject effects. There was a significant difference in blaming different subjects, $F_{(3.27, 2135.64)}$ = 354.15, p < .001, η_p^2 = .352, as well as a significant effect of the scenarios, $F_{(3, 653)}$ = 4.72, p < .003, η_p^2 = .021. However, these main effects were qualified by a significant interaction between scenarios and the blame of different entities, $F_{(9.81, 2135.64)}$ = 5.78, p < .001, η_p^2 = .026. There was neither a significant difference in blame of the operator between the scenarios, $F_{(3, 653)}$ = 1.83, p = .141, nor a difference in blame of the commanding officer, $F_{(3, 653)}$ = 0.573, p = .633. On the other hand, the blame of the programmer of the targeting mechanism, $F_{(3, 653)}$ = 3.67, p = .012, η^2 = .017, the company manufacturing the machine, $F_{(3, 653)}$ = 7.53, p < .001, η^2 = .033, and the machine itself, $F_{(3, 653)}$ = 14.0, p < .001, η^2 = .060, differed between the scenarios. Specifically, the blame for all these three subjects grew with the increasing level of autonomy in the scenarios ($t_{\text{linear contrast}}$s from 2.94 to 6.42, all ps < .003). Age was negatively related to blaming of all entities across all scenarios. Positive attitudes toward the military were significantly related to lesser blaming of all entities across all scenarios, Fs $_{(1, 648)}$ from 10.44 to 41.81, ps < .001, with the exception of blaming the machine, in which case the effect was not significant, $F_{(1, 648)}$ = 0.183, p = .669. The other variable significantly related to the amount of blame was political identification: higher conservatism was related to higher blame of the machine across all scenarios, F (1, 648) = 15.893, p < .001, η2 = .024, and to lower blame of the commanding officer, regardless of the scenario, F (1, 648) = 17.71, p < .001, η2 = .027.

Analysis 2 – ethicality and approval of the strike

We used ANCOVAs to analyze whether the participants found the strike ethical and how much they approved of it, while controlling for the level of education, income, political identification, attitudes toward the military, and gender. There was a significant difference in ethicality judgments between the scenarios, $F_{(3, 647)} = 3.406$, $p = .017$, $\eta^2 = .016$ – the strikes in scenarios with increasing drone autonomy were judged as less ethical ($t = 2.60$, $p = .009$). Gender also had a main effect, $F_{(1, 647)} = 5.837$, $p = .016$, $\eta^2 = .009$, with male participants judging the strike as less ethical; attitude toward the military had a main effect, $F_{(1, 648)} = 112.219$, $p < .001$, $\eta^2 = .148$; as did political identification, $F_{(1, 647)} = 59.158$, $p < .001$, $\eta^2 = .084$. The results of the similar analysis for approval of the strike were in the same direction, but the difference between scenarios was not significant when tested as a between subject effect, $F_{(1, 647)} = 1.624$, $p < .183$, $\eta^2 = .007$. However, because we asked participants to rate approval of the strike in all scenarios at the end of the survey, we could analyze the difference between scenarios using a repeated measures ANOVA. Once again, we applied the Greenhouse-Geisser correction ($\varepsilon = .874$) based on the significant test of sphericity, Macuhly's $W = .808$, p < .001. Based on this analysis with a higher statistical power, the approval of the strike decreased with the increasing autonomy, $F_{(2.62, 2636.30)} = 77.5$, $p < .001$, $\eta^2 = .07$.

Analysis 3 – meaningful human control

Most experts at the UN agree on a requirement to retain some degree of human involvement, or “meaningful human control” (MHC), over the target selection and engagement functions of weapon systems. One of the key purposes behind the emerging MHC norm is ensuring that somebody can be held responsible for potential rule-violations resulting from AWS use, and thereby avoiding the “responsibility gaps” (Horowitz & Scharre, 2015). There is currently no universal definition of MHC (cf. Amoroso & Tamburrini, 2019; Horowitz & Scharre, 2015; Roff & Moyes, 2016). Nevertheless, drawing on the commonalities between different existing conceptualizations (see Ekelhof, 2019), we can identify at least some minimum requirements for weapons to be considered operating under meaningful human control. These include:

1. There are means available to the commander/operator to intervene and, potentially, abort the attack.
2. Commanders/operators are conscious of all decisions to use force they make, approve, or monitor.
3. Commanders/operators have enough time and information to consider the context of the situation.

The above-identified elements were used to formulate Scenarios 4 and 5, which specify whether the system complied with (Scenario 4) or violated (Scenario 5) the MHC norm. We hypothesized that compliance with MHC requirements would increase the public approval of the strike and the perception of its ethicality. Furthermore, we expected that compliance with MHC requirements would further shift the blame toward human agents (i.e., the commander, the operator, and the programmer). Conversely, informing the respondents that the MHC norm was violated should lead to the opposite trend.

After collecting the data, we repeated the analyses described in the main text, comparing the human-supervised autonomous drone scenario not mentioning the elements of MHC with the scenarios in which MHC was either explicitly upheld or violated (see Figure 3). For an analysis of the blame, we used a 2-factor mixed ANOVA. There was a significant interaction between scenarios and blame of different entities, $F_{(6.67,\ 1701.27)} = 2.44$, $p = .019$. We therefore analyzed the blame levels for each entity separately. There was no significant difference in blame of the drone operator, commanding officer, programmer, or machine between the scenarios. On the other hand, the blame of the company manufacturing the machine was higher when the MHC was violated than when it was upheld, $t_{Tukey}(510) = 3.114$, $p = .006$.

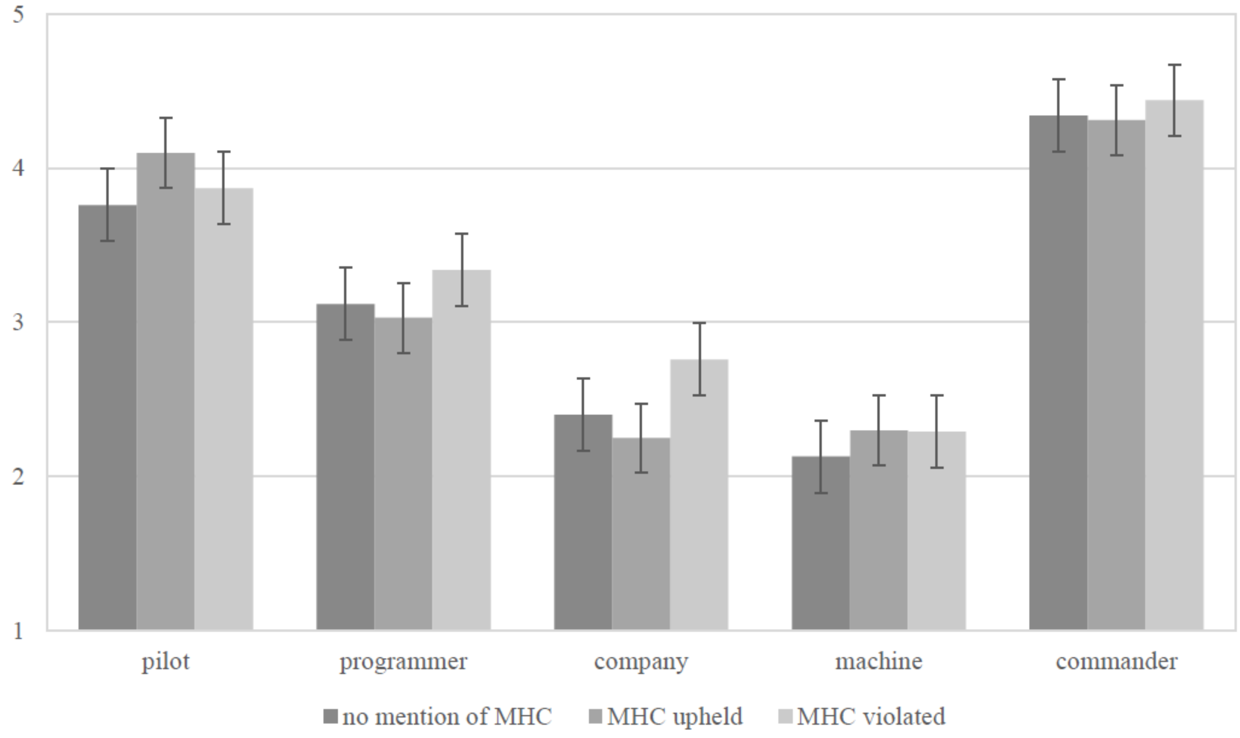

Figure 3: Blame Attribution and the MHC

Note. All blame judgments were made on 6-point scales ranging from 1 – no blame to 6 – maximum blame. Error bars represent 95% confidence intervals.

Finally, we analyzed the ethicality and approval of the strike based on the MHC (see Figure 4). The strike in the scenario in which MHC was upheld was judged as significantly more ethical than the strike in the scenario in which the MHC was violated, ($t_{Tukey}(510) = 2.43$, p

= .041). The results for the approval were similar – again, the strike in the case of the upheld MHC was approved more than when MHC was not mentioned ($t_{Tukey}(510) = 2.83$, $p = .014$) or when it was violated ($t_{Tukey}(510) = 2.80$, $p = .015$). When participants judged all scenarios, the difference in approval between upholding and violating MHC remained significant, $t_{Tukey}(2008) = 2.92$, $p = .010$, however, the average approval in the scenario where MHC was upheld decreased substantially (see Figure 4).

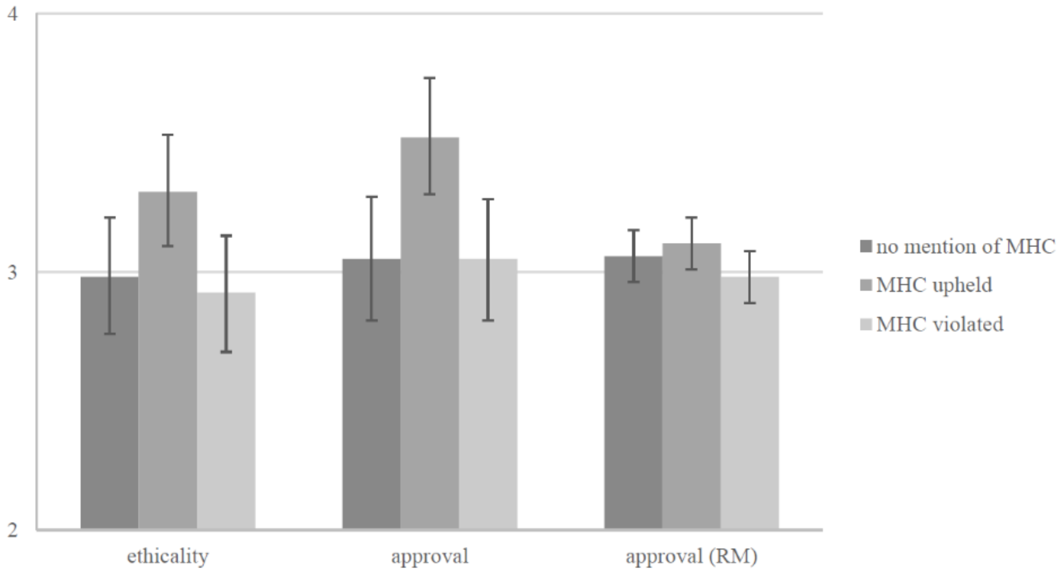

Figure 4: MHC and the judgment of the ethicality and approval of the strike

Note. The “ethicality” and “approval” judgments are based on evaluations of a single scenario presented at the beginning of the survey. The “approval (RM)” judgments are based on parallel evaluations of all scenarios made by participants at the end of the survey. All judgments were made on 6-point scales ranging from 1 – strongly disapprove / highly unethical to 6 – strongly approve / highly ethical. Error bars represent 95% confidence intervals.

The results show that specifying whether a system operates under MHC makes little difference for blame attribution. The only exception was a slight increase in the blame of the company that manufactured the machine in case of the MHC violation. Furthermore, while specifying that MHC requirements were upheld led to a statistically significant increase in the perceived ethicality and approval of the strike, this increase was relatively low and especially low when participants evaluated all scenarios together (RM).

References

Amoroso, D., & Tamburrini, G. (2019). What makes human control over weapons systems "meaningful"? *International Committee for Robot Arms Control*. https://www.icrac.net/wp-content/uploads/2019/08/Amoroso-Tamburrini_Human-Control_ICRAC-WP4.pdf

Dupuis, M., Endicott-Popovsky, B., & Crossler, R. (2013). An Analysis of the Use of Amazon's Mechanical Turk for Survey Research in the Cloud. *Proceedings of the International Conference on Cloud Security Management*, October 2013, 17–18. http://faculty.washington.edu/marcjd/articles/An%20Analysis%20of%20the%20Use%20of%20Amazon%E2%80%99s%20Mechanical%20Turk%20for%20Survey%20Research%20in%20the%20Cloud%20-%20v.%203.pdf

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

Horowitz, M. C., & Scharre, P. (2015). Meaningful human control in weapon systems: A primer. *Center for a New American Security*. https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/Ethical_Autonomy_Working_Paper_031315.pdf?mtime=20160906082316&focal=none

Roff, H., & Moyes, R. (2016). Meaningful human control, artificial intelligence and autonomous weapons. *Article 36*. https://article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf