ARTICLE

THE CONCEPT OF "THE HUMAN" IN THE CRITIQUE OF AUTONOMOUS WEAPONS

Kevin Jon Heller[*]

ABSTRACT

The idea that using "killer robots" in armed conflict is unacceptable because they are not human is at the heart of nearly every critique of autonomous weapons. Some of those critiques are deontological, such as the claim that the decision to use lethal force requires a combatant to suffer psychologically and risk sacrifice, which is impossible for machines. Other critiques are consequentialist, such as the claim that autonomous weapons will never be able to comply with international humanitarian law (IHL) because machines lack human understanding and the ability to feel compassion.

This article challenges anthropocentric critiques of AWS. Such critiques, whether deontological or consequentialist, are uniformly based on a very specific concept of "the human" who goes to war: namely, someone who perceives the world accurately, understands rationally, is impervious to negative emotions, and reliably translates thought into action. That idealized individual, however, does not exist; decades of psychological research make clear that cognitive and social biases, negative emotions, and physiological limitations profoundly distort human decision-making—particularly when humans find themselves in dangerous and uncertain situations like combat. Given those flaws, and in light of rapid improvement in sensor and AI technology, it is only a matter of time until autonomous weapons are able to comply with IHL better than human soldiers ever have or ever will.

CONTENTS

[*] Professor of International Law & Security, Department of Political Science, University of Copenhagen (Centre for Military Studies); Special Advisor to the Prosecutor of the International Criminal Court on War Crimes.

INTRODUCTION

Autonomous weapons—“robotic weapon systems that, once activated, can select and engage targets without further intervention by a human operator”[1]—are no longer the stuff of science fiction. Not only have weapons with a fully autonomous mode[2] been used for offensive purposes[3] in Libya, Syria, and Nagorno-Karabakh in the last three years alone, both sides of the conflict in Ukraine have used such weapons: Russia uses Lancet drones, which circle a predetermined geographic area and then engage a preselected target without human intervention,[4] and Ukraine uses Punisher drones, which can attack a target without human intervention when used in tandem with a smaller reconnaissance drone.[5] It is only

[1] Christof Heyns (Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions), *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, ¶ 38, U.N. Doc. A/HRC/23/47 (Apr. 9, 2013). As the U.S. Department of Defense notes, “[t]his includes, but is not limited to, operator-supervised autonomous weapon systems that are designed to allow operators to override operation of the weapon system, but can select and engage targets without further operator input after activation.” U.S. DEP’T OF DEF., DIR. 3000.09, AUTONOMY IN WEAPONS SYSTEMS, para. G.2. (Jan. 25, 2023).

[2] It is unclear whether Lancets or Punishers have been used fully autonomously because both types of drone are normally used with a human supervising the targeting missions. *See, e.g.,* Jeremy Kahn, *A.I. is on the Front Lines of the War in Ukraine*, FORTUNE (Mar. 1, 2022), https://fortune.com/2022/03/01/russia-ukraine-invasion-war-a-i-artificial-intelligence [https://perma.cc/X9Q7-E8T7].

[3] Autonomous weapons, such as the Phalanx close-in weapons system, have long been used for defensive purposes. *See generally* VINCENT BOULANIN & MAAIKE VERBRUGGEN, MAPPING THE DEVELOPMENT OF AUTONOMY IN WEAPON SYSTEMS (2017).

[4] Kahn, *supra* note 2.

[5] *See* Haye Kesteloo, *Punisher Drones Are Positively Game-Changing for Ukrainian Military in Fight Against Russia*, DRONEXL (Mar. 3, 2022), https://dronexl.co/2022/03/03/punisher-drones-ukrainian-military [https://perma.cc/94EZ-7CCU].

a matter of time before autonomous weapons systems (AWS) become ubiquitous on the battlefield,[6] because the military advantages they potentially provide, particularly targeting precision, speed, and force protection, have led to an increasingly frantic race to develop them. The United States spent $18 billion on unmanned-systems research between 2016 and 2020,[7] while Russia has declared its intention to have one-third of its military run by artificial intelligence (AI) no later than 2030.[8] As Russian President Vladimir Putin said in 2017, "[w]hoever becomes the leader" in AI will "become the ruler of the world."[9]

The urgency with which states are developing autonomous weapons has been met by equally urgent efforts to ban them. The Campaign to Stop Killer Robots was founded in 2013 and is currently supported by more than 200 civil society organizations worldwide, including Human Rights Watch and Amnesty International.[10] In the Campaign's view, it is unacceptable for non-human machines to take human life regardless of whatever military advantages they might provide:

> Autonomy in weapons systems is a profoundly human problem. Killer robots change the relationship between people and technology by handing over life and death decision-making to machines. They challenge human control over the use of force, and where they target people, they dehumanise us—reducing us to data points. But technologies are designed and created by people.[11]

The Campaign is thus calling on states to negotiate an international treaty that would prohibit "autonomous weapons systems that do not allow for meaningful

[6] *See, e.g.*, Michael N. Schmitt & Jeffrey S. Thurnher, *"Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict*, 4 HARV. NAT'L SEC. J. 231, 237 (2013) (noting that some Department of Defense studies "have even suggested that autonomous weapons may become the norm on the battlefield in a generation"); Kenneth Anderson & Matthew C. Waxman, *Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can* 27 (Jean Perkins Task Force on National Security and Law Essay Series, 2013), http://www.ssrn.com/abstract=2250126 [https://perma.cc/F76E-DRTC] ("[I]ncremental development and deployment of autonomous weapon systems is inevitable.").

[7] James Dawes, *What You Need to Consider About "Killer Robots" and Autonomous Weapons Research*, FAST CO. (Dec. 29, 2021), https://www.fastcompany.com/90707966/what-you-need-to-consider-about-killer-robots-and-autonomous-weapons-research [https://perma.cc/7ASR-3QSD].

[8] David Freedman, *US Is Only Nation with Ethical Standards for AI Weapons. Should We Be Afraid?*, NEWSWEEK (Sept. 15, 2021), https://www.newsweek.com/2021/09/24/us-only-nation-ethical-standards-ai-weapons-should-we-afraid-1628986.html [https://perma.cc/EZ8R-XZPB].

[9] James Vincent, *Putin Says the Nation That Leads in AI "Will Be the Ruler of the World,"* THE VERGE (Sept. 4, 2017), https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world [https://perma.cc/T7G4-3U9A].

[10] *About Us*, STOP KILLER ROBOTS, https://www.stopkillerrobots.org/about-us [https://perma.cc/9F39-VWCZ] (last visited Oct. 3, 2023).

[11] *Problems with Autonomous Weapons*, STOP KILLER ROBOTS, https://www.stopkillerrobots.org/stop-killer-robots/facts-about-autonomous-weapons [https://perma.cc/SD2S-XLMJ] (last visited Oct. 3, 2023).

human control."[12] To date, at least seventy states have heeded its call, including the United States and the United Kingdom.[13]

There is nothing idiosyncratic about the Campaign's focus on the inhumanity of autonomous weapons. On the contrary, the idea that using "killer robots" in armed conflict is unacceptable because they are not human is at the heart of nearly every critique of AWS. Some of those critiques are deontological,[14] such as the claim that the decision to use lethal force requires a combatant to suffer psychologically and risk sacrifice, which is impossible for machines. Other critiques are consequentialist,[15] such as the claim that autonomous weapons will cause unnecessary death in conflict because their lack of human understanding and inability to feel compassion make them incapable of complying with international humanitarian law (IHL).

This article challenges critiques of autonomous weapons that focus on their lack of humanity. Anthropocentric critiques of AWS implicitly contrast machines with a very specific concept of "the human" who goes to war: namely, someone who is generally capable of perceiving the world accurately, understanding rationally, quarantining negative emotions, and reliably translating thought into action. That idealized individual, however, does not exist: decades of psychological research make clear that cognitive and social biases, negative emotions, and physiological limitations profoundly distort human decision-making—particularly when humans find themselves in dangerous and uncertain situations like combat. It is precisely humans as they are, not as critics of autonomous weapons imagine them to be, that explains why combat using machines will eventually be more ethical and more humane than combat with human soldiers.

The article is divided into five sections. Section I critiques deontological objections to autonomous weapons. It shows that those objections wrongly anthropomorphize AWS by assuming they "decide" on targets in a manner similar to humans, overstate the inclination of humans to think about the consequences of killing and to risk sacrificing themselves for others, and are predicated on a

[12] *Our Policy Position*, STOP KILLER ROBOTS, https://www.stopkillerrobots.org/our-policies [https://perma.cc/GLF8-MM4Z] (last visited Oct. 3, 2023).

[13] *70 States Deliver Joint Statement on Autonomous Weapons Systems at UN General Assembly*, STOP KILLER ROBOTS (Oct. 21, 2022), https://www.stopkillerrobots.org/news/70-states-deliver-joint-statement-on-autonomous-weapons-systems-at-un-general-assembly [https://perma.cc/3D3U-XH2J]. As the statement notes, however, states disagree over what "meaningful human control" requires. *Id.*

[14] A deontological critique is one "that would count against the use of AWS even if AWS were to yield optimal outcomes vis-à-vis our legitimate military aims." Michael Skerker, Duncan Purves & Ryan Jenkins, *Autonomous Weapons Systems and the Moral Equality of Combatants*, 22 ETHICS INF. TECHNOL. 197, 198 (2020).

[15] A consequentialist objection is one that assesses "moral obligations and permissions exclusively on the basis of an evaluation of the (actual or expected) consequences of actions." Daniele Amoroso & Guglielmo Tamburrini, *The Ethical and Legal Case Against Autonomy in Weapons Systems*, 18 GLOB. JURIST 1, 10 (2017).

romanticized and anachronistic view of war in which most killing takes place face-to-face.

Sections II and III critique consequentialist objections to autonomous weapons that focus on *jus in bello*. Section II, the longest section of the article, addresses the common argument that IHL compliance requires human understanding—particularly the ability to discern the intentions of potential targets and to make fact-sensitive and context-dependent determinations. The section begins by demonstrating that such understanding is far less necessary to IHL than AWS critics assume. It then explains why, in those situations where understanding is necessary, well-documented limits on human decision-making undermine the idea that human soldiers are more likely to comply with IHL than autonomous weapons. Finally, the section ends by discussing why the concept of "meaningful human control" is an undesirable solution to the supposed problems of AWS and should give way to the superior concept of "meaningful human certification."

Section III responds to the claim that autonomous weapons should be prohibited because machines cannot feel compassion, an emotion that is both ethically and legally required on the battlefield. It makes three arguments. The first is that compassion is irrelevant to IHL compliance. The second is that the potential benefits of compassion in combat are far outweighed by the costs of negative emotions such as stress and anger. The third is that compassion can lead to negative outcomes in combat as well as positive ones.

Section IV focuses on international criminal law, addressing the argument that the non-human nature of autonomous weapons makes it difficult, if not impossible, to hold humans responsible for war crimes AWS may commit. The section shows not only that the problem of "accountability gaps" is significantly overstated, but also that there is no significant difference between human soldiers and autonomous weapons in terms of criminal responsibility.

Finally, Section V explores a consequentialist objection to autonomous weapons that focuses on the *jus ad bellum*: namely, that replacing human soldiers with non-human machines will reduce the number of casualties during an armed conflict, making it easier for democratic states to go to war. The section argues that this is the most persuasive objection to AWS—and one that is actually understated, because it ignores the potential for such weapons to minimize civilian casualties, another factor that affects a state's willingness to use armed force. As the section notes, however, the *jus ad bellum* critique is less an objection to AWS than to modern warfare itself, because most of the weapons developed over the past century have had precisely the same effect.

I. THE HUMAN AS BONUM IN SE

Critics often assert that autonomous weapons are "*mala in se*"[16]—evil in themselves, regardless of the consequences of their use. Four such deontological arguments, all of which are interrelated, are particularly common. Three focus primarily on the individual who kills, claiming that machine killing is inherently wrong because taking life requires the killer to possess morality, to suffer psychologically, and to risk sacrifice. The fourth, by contrast, focuses primarily on the individual who is killed, insisting that machine killing necessarily dehumanizes the victim.

A. Only Humans Have Morality

The first deontological objection to autonomous weapons is that it is unethical for machines to kill because the decision to take life is so profound that it requires the kind of moral judgment only humans possess. In the words of Christof Heyns, the former United Nations Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions:

> A machine, which is bloodless and without morality or mortality, cannot do justice to the gravity of the decision whether to use force in a particular case, even if it may be more accurate than humans. This decision is so far-reaching that each instance calling for its use requires that a human being should decide afresh whether to cross that threshold if it is not to become a mechanical—and inhuman—process.[17]

Many other scholars emphasize what they view as the necessary connection between human morality and the ethical use of lethal force. Daniele Amoroso et al., for example, claim that "[i]n order to be non-arbitrary. . . the act of killing must be grounded on human judgement, for only human decision-making guarantees the full appreciation of the value of individual life and the significance of its loss."[18] Similarly, Srđan Korać argues that "[t]he decision to kill in the context of military operations must remain exclusively an act of human free will as the characteristic

[16] *See, e.g.*, Robert Sparrow, *Robots and Respect: Assessing the Case Against Autonomous Weapon Systems*, 30 ETHICS INT. AFF. 93, 100 (2016) ("AWS should be acknowledged as *mala in se* by virtue of the extent to which they violate the requirement of respect for the humanity of our enemies, which underlies the principles of *jus in bello*.").

[17] Christof Heyns, *Autonomous Weapons Systems: Living a Dignified Life and Dying a Dignified Death*, *in* AUTONOMOUS WEAPONS SYSTEMS: LAW, ETHICS, POLICY 3, 11 (Nehal Bhuta et al. eds., 2016).

[18] DANIELE AMOROSO ET AL., AUTONOMY IN WEAPON SYSTEMS: THE MILITARY APPLICATION OF ARTIFICIAL INTELLIGENCE AS A LITMUS TEST FOR GERMANY´S NEW FOREIGN AND SECURITY POLICY 32 (2018).

inherent to human conscience—the only possible basis for ethical reasoning about whether an action can be evaluated as right or wrong."[19]

As the three quotes indicate, the "moral judgment" objection to machine killing necessarily assumes that an autonomous weapon "decides" to take human life in a manner akin to human decision-making. In Armin Krishnan's words, "[w]hat is scary about the killer robot is not the fact that it would be more dangerous than mines, but rather its ability to make life and death decisions in place of a human."[20] Indeed, for Krishnan, the fact that an autonomous weapon is capable of deciding to kill "elevates it ontologically and maybe even morally from the mere object to a subject capable of morally meaningful action."[21]

The problem with this argument is that—ironically—it anthropomorphizes autonomous weapons. Killer robots do not decide whom to target in a manner akin to human soldiers. Indeed, they do not "decide" at all. They simply execute the targeting rules that humans have programmed into them prior to their activation:

> While such systems may not be programmed with precise predetermined responses to every situation they encounter, they are programmed with some system for developing a response and are thereby operating in accordance with their programming regardless of whether or not the specific behaviours they in fact adopt were or could have been foreseen during development or at the time of deployment. Such a machine is still just an instrument of the will of its developers and those responsible for employing it in some

[19] Srđan Korać, *Depersonalisation of Killing: Towards a 21st Century Use of Force "Beyond Good and Evil?",* 29 FILOZ DRUS 49, 162 (2018); *see also* Elvira Rosert & Frank Sauer, *Prohibiting Autonomous Weapons: Put Human Dignity First*, 10 GLOB. POL'Y 370, 370 (2019) ("The minimum requirement for upholding human dignity, even in conflicts, is that life and death decisions on the battlefield should always and in principle be made by humans.").

[20] ARMIN KRISHNAN, KILLER ROBOTS: LEGALITY AND ETHICALITY OF AUTONOMOUS WEAPONS 33 (2010); *see also* Amoroso & Tamburrini, *supra* note 15, at 8 (claiming that "AWS taking human life are neither legitimate nor morally justifiable" because "lethal decisionmaking is carried out by a machine without any involvement of human judgement"); Taylor Jones, *An Introduction to the Issue of Lethal Autonomous Weapons*, FUTURE OF LIFE INSTITUTE (Nov. 30, 2021) https://futureoflife.org/aws/an-introduction-to-the-issue-of-lethal-autonomous-weapons [https://perma.cc/8SUB-AN7J] ("Algorithms are incapable of comprehending the value of human life, and so should never be empowered to decide who lives and who dies."); *cf.* Duncan Purves, Ryan Jenkins & Bradley J. Strawser, *Autonomous Machines, Moral Judgment, and Acting for the Right Reasons*, 18 ETHICAL THEORY MORAL PRAC. 851, 866 (2015) ("Surely an AWS is not totally inert; its purpose is precisely to make decisions about who should live or die; to discriminate on its own between targets and courses of action; indeed, to fulfill all of the purposes that a soldier would fulfill in its place.").

[21] KRISHNAN, *supra* note 20, at 33; *see also* HUMAN RIGHTS WATCH, MIND THE GAP: THE LACK OF ACCOUNTABILITY FOR KILLER ROBOTS 6 (2015) (arguing that, "for some legal analyses, [an AWS] would be more akin to a human soldier than to an inanimate weapon").

> situation; it is not accurately characterized as an independent decisionmaker.[22]

This is not a semantic distinction. Because an autonomous weapon's selection of targets is always a function of its programming, it makes no sense to object to AWS on the ground that only humans can ethically decide to take human life. When an autonomous weapon uses lethal force, a human *has* decided to kill: namely, the human—or humans—responsible for determining which kinds of individuals and objects the machine will target.[23] Indeed, even a staunch critic like Robert Sparrow admits as much.[24]

To be sure, autonomous targeting is different than non-autonomous targeting. In most situations, a human uses a non-autonomous weapon—whether a rifle or a "fire and forget" precision-guided munition[25]—to engage a specific target or specific group of targets. By contrast, "[i]n the case of an autonomous weapon . . . the human has decided to launch a weapon to seek out and destroy a general class of targets over a wide area but is not making a decision about which specific targets are to be engaged."[26] It is precisely this aspect of autonomous targeting that creates the appearance of an AWS acting with human-like free will: because its targeting rules are general—expressed, for example, in the form "if <camera image matches image in database with probability of more than 95%> then <aim and fire> else

[22] TIM MCFARLAND, AUTONOMOUS WEAPON SYSTEMS AND THE LAW OF ARMED CONFLICT: COMPATIBILITY WITH INTERNATIONAL HUMANITARIAN LAW 1328 (2020); Rebecca Crootof, *The Killer Robots Are Here: Legal and Policy Implications*, 36 CARDOZO L. REV. 1837, 1848 (2015) ("[A]utonomous weapon systems will also operate under an array of preprogrammed and practical constraints; an autonomous weapon system will not simply be directed to 'eliminate the enemy.'"); DEANE-PETER BAKER, SHOULD WE BAN KILLER ROBOTS? 87–88 (2022) ("To speak of LAWS as making the decision to kill is to anthropomorphize LAWS and at the same time to play fast and loose with language. And if a LAWS has not made the decision to kill, then it is nonsensical to speak of the human operator as having somehow 'ceded' or 'delegated' that decision to the LAWS."); BOULANIN & VERBRUGGEN, *supra* note 3, at 10 (noting that although AWS "do not simply go through a series of pre-scripted actions," that does not mean "their behaviour is not predictable or that the systems are capable of free will. Control systems do only what they are programmed to do, regardless of the complexity of their programming"); Eliav Lieblich & Eyal Benvenisti, *The Obligation to Exercise Discretion in Warfare: Why Autonomous Weapons Systems Are Unlawful*, *in* AUTONOMOUS WEAPONS SYSTEMS: LAW, ETHICS, POLICY 245, 250 (Nehal Bhuta et al. eds., 2016) ("Since a technically autonomous machine is incapable of altering its algorithms through a process that could be equated to human learning, we cannot claim that it engages in a true process of discretion—its 'inner deliberation' is controlled by determinations made *ex ante* by humans.").

[23] *See, e.g.*, Skerker, Purves & Jenkins, *supra* note 14, at 199 (noting that "it is questionable whether an AWS can sever the interpersonal relationship between the target of the AWS and the person who deploys the AWS unless AWS are full-blooded moral agents").

[24] Sparrow, *supra* note 16, at 107 ("We might equally well think of a robot as a tool by which one person attempts to kill another—albeit an indeterminate other. The relevant interpersonal relationship would then be that between the officer who authorizes the release of the weapon and those the officer intends to kill.").

[25] A "fire and forget" munition is one that cannot be recalled after it is launched. PAUL SCHARRE & MICHAEL C. HOROWITZ, AN INTRODUCTION TO AUTONOMY IN WEAPON SYSTEMS 9 (2016).

[26] *Id*. at 16.

<keep searching>"[27]—it is impossible to know *ex ante* each and every target it will engage. But that does not mean the autonomous weapon is "a subject capable of morally meaningful action."[28] On the contrary, even the most sophisticated AWS engaging in the most open-ended targeting remains a mechanism for giving effect to its operator's intentions.[29]

Seemingly recognizing this fact, the International Committee of the Red Cross (ICRC) offers a slightly different moral-judgment critique of autonomous weapons. In its view, taking life cannot be ethical unless the intent of the person initiating the attack is "directly linked" to its "eventual outcome."[30] The problem with killer robots, according to the ICRC, is that they sever the connection between intent and result:

> [T]he key issue is that the commander or operator activating the weapon is not giving instructions on a specific target to be attacked ("specific armoured vehicle") at a specific place ("at the corner of that street") and at a specific point in time ("now"). Rather, when activating the autonomous weapon system, by definition, the user will not know exactly which target will be attacked ("armoured vehicles fitting this technical signature"), in which place (within x square kilometres) or at which point in time (during the next x minutes/hours). Thus, it can be argued, this more generalized nature of the targeting decision means the user is not applying their intent to each specific attack.[31]

The problem with this argument is that it applies equally to a variety of weapons and military tactics that have long been deployed without ethical objection.[32] One example of the former is the NATO-standard Mark 46 acoustic

[27] MCFARLAND, *supra* note 22, at 1327.

[28] *See, e.g.*, SCHARRE AND HOROWITZ, *supra* note 25, at 16 (noting that the definitions of autonomous and semi-autonomous weapons each "focus on the decision the human is making or not making and do not apply the word 'decision' to something the weapon itself is doing").

[29] *See, e.g.*, Magdalena Pacholska, *Military Artificial Intelligence and the Principle of Distinction: A State Responsibility Perspective*, ISR. L. REV. 1, 8–9 (2022) ("Even the most advanced versions of AWS . . . 'select' specific targets from a human pre-defined class or category."); Sparrow, *supra* note 16, at 107 ("Neither the fact that the person who authorizes the launch does not know precisely who she is killing when she sends an AWS into action nor the fact that the identity of those persons may be objectively indeterminate at the point of launch, seems to rule out the possibility of the appropriate sort of relationship of respect."); BAKER, *supra* note 22, at 59 ("In everyday language the claim here is that the prerequisites for an agent to have control over something are that the agent (a) has intentions about the state or behaviour of that thing, and (b) the ability to cause the thing in question to go into the intended state or to behave in the intended way.").

[30] INTERNATIONAL COMMITTEE OF THE RED CROSS, ETHICS AND AUTONOMOUS WEAPON SYSTEMS: AN ETHICAL BASIS FOR HUMAN CONTROL? 11 (2018).

[31] *Id.* at 12.

[32] *See* BAKER, *supra* note 22, at 45–46 ("[T]he epistemic conditions we set as requirements for the ethical trustworthiness of LAWS must be reasonable. If . . . a degree of generalization in targeting

homing torpedo, which has been in service since the early 1960s. When fired from a helicopter in response to a sonar contact, the Mark 46 torpedo seeks out any submarine within its range—not "a particular submarine at a particular time and in a particular location."[33] Similarly, once activated, the CAPTOR deep-water mine—employed during the Cold War—used its ability to distinguish between friendly and enemy acoustic signatures to automatically fire its torpedo at any Soviet submarine that came within range of its sonar system.[34]

The ICRC's intent requirement would also prohibit a number of firing tactics that have been central to ground warfare for decades, if not centuries. "Reconnaissance by fire" involves placing fire "on a suspected enemy position to cause the enemy to disclose his presence by movement or return fire."[35] As soon as the enemy comes into range, "long range fire" is designed "to engage the enemy as early as possible to inflict casualties, delay his advance, harass him, interdict him, and disrupt his organization."[36] "Barrage fire" is intended to "fill a volume of space or area rather than aimed specifically at a given target."[37] "Distributed fire" is fire "so dispersed as to engage most effectively an area target"[38]—including areas "in which the exact location of the enemy is unknown."[39] None of these tactics involve soldiers intending to attack a specific target or specific group of targets. Their intent is instead to attack a particular geographic space in which any individual or object struck by fire is presumed to be a legitimate military objective.

There is, in short, no qualitative difference between autonomous and non-autonomous weapons in terms of the kinds of targets they attack. Both can be used to attack specific targets, specific groups of targets, or general classes of targets. But that does not mean there are no important differences between the two. On the contrary, they differ significantly in terms of *accuracy*: autonomous weapons have the potential to target with far more precision than even the most precise non-autonomous weapons—especially those that rely on the skill of their human

is appropriate for artillery batteries and for homing torpedoes, then counting LAWS as ethically problematic because their employment means giving up 'specificity' is equally puzzling.").

[33] *Id*. at 42–43.

[34] ROBERT O. WORK, PRINCIPLES FOR THE COMBAT EMPLOYMENT OF WEAPON SYSTEMS WITH AUTONOMOUS FUNCTIONALITIES 6 (2021).

[35] U.S. DEP'T OF ARMY, TC 7-100.2, OPPOSING FORCE TACTICS para. 8-29 (Dec. 9, 2011), https://odin.tradoc.army.mil/TC/TC_7-100.2_Opposing_Force_Tactics/TC_7-100.2_Opposing_Force_Tactics [https://perma.cc/JF2L-VZWQ].

[36] U.S. DEP'T OF ARMY, TC 3-22.91 MORTAR FIRE DIRECTIONS AND PROCEDURES para. 3-103 (May 15, 2017), https://armypubs.army.mil/epubs/DR_pubs/DR_a/pdf/web/ARN3488_TC%203-22x91%20FINAL%20WEB%201.pdf [https://perma.cc/Q7TG-Q3LN].

[37] *Barrage Fire*, NATO TERMINOLOGY DATABASE, https://nso.nato.int/natoterm/Web.mvc [https://perma.cc/9GDR-45CS] (last visited Oct. 9, 2023).

[38] *Distributed Fire*, NATO TERMINOLOGY DATABASE, https://nso.nato.int/natoterm/Web.mvc [https://perma.cc/7LLS-ZPXP] (last visited Oct. 9, 2023).

[39] U.S. DEP'T OF ARMY, FIELD MANUAL 3-22.68, COMBAT TECHNIQUES OF FIRE, para. 5–9 (Jan. 31, 2003), https://www.globalsecurity.org/military/library/policy/army/fm/3-22-68/c05.htm [https://perma.cc/4JVN-XKTS].

operator, such as a rifle or a mortar.[40] As former United States Deputy Secretary of Defense Robert Work says, "improved autonomous functionalities in navigation, target identification, and mid-course and terminal guidance have led to a wholesale shift to guided weapons that are far more accurate than previous generations of unguided weapons."[41] Indeed, even AWS's most ardent critics accept that they hold out the prospect of unprecedented targeting accuracy.[42] The autonomy of autonomous weapons is thus a positive, not a negative, because it is precisely their non-human nature—their lack of dependence on human ability—that makes such accurate targeting possible.

B. Only Humans Suffer

The second deontological objection to autonomous weapons is that machines cannot ethically kill because taking life requires the uniquely human ability to suffer the psychological consequences. Linda Eggert claims, for example, that "AWS do not, in any relevant sense, struggle with obstacles in a way that would make it so difficult to do the right thing that they should be legally permitted to do what is morally wrong."[43] Similarly, Elvira Rosert and Frank Sauer insist that "fundamental humanitarian norms" are lost when a military "outsourc[es] moral costs by no longer concerning itself with the act of killing, with no individual combatants' psyches burdened by the accompanying responsibility."[44]

One problem with this objection is that, in practice, very few human soldiers in a firefight actually contemplate the implications of taking life, much less have "psyches burdened by the accompanying responsibilities." As Daniel Lim says, the battlefield "is arguably the worst possible environment to expect something along these lines" because human soldiers "hardly have the time or mental/emotional space to exercise the concept of sacrifice or generate the relevant emotions to make informed decisions each time they deploy lethal force."[45] On the contrary, in the heat of battle, most soldiers simply want to survive.[46]

[40] *See, e.g.*, MCFARLAND, *supra* note 22, at 178 ("Precision through superior sensing, superior control over the quantum of force applied to a target and other advanced capabilities are among the benefits promised by further AWS development."); BAKER, *supra* note 22, at 44 ("Rather than being limited to saturating a specific geographical area with high explosive munitions, employing a well-designed LAWS potentially allows for the addition of further parameters, which give greater opportunities for specificity in the application of the targeteer's intent.").

[41] WORK, *supra* note 34, at 8.

[42] *See, e.g.*, Amoroso & Tamburrini, *supra* note 15, at 12 (acknowledging that "consequentialist arguments may lead one to support the introduction of AWS for targeting humans in view of their greater targeting accuracy"); Heyns*, supra* note 17, at 7 ("Automation of force can arguably allow greater speed and accuracy in targeting or preventing the excessive use of force.").

[43] LINDA EGGERT, AUTONOMOUS WEAPONS AND WHY THE LAWS OF WAR ARE NOT ENOUGH 6 (2022).

[44] Rosert & Sauer, *supra* note 19, at 373.

[45] DANIEL LIM, KILLER ROBOTS AND HUMAN DIGNITY 174 (2019).

[46] *See, e.g.*, Anzhelika Solovyeva & Nik Hynek, *Going Beyond the "Killer Robots" Debate*, 12 CENT. EUR. J. INT'L & SEC. STUD. 166, 182 (2018) (noting that "in real combat, only a few combatants may seek combat glory, while roughly ninety-nine percent of them simply want to

To be sure, soldiers who kill to live may have burdened psyches after a firefight is over. That is a relevant difference between human soldiers and autonomous weapons. But not all soldiers feel the emotional burden of taking life *ex post*—particularly those who kill remotely. Numerous empirical studies have found, for example, that unmanned aerial vehicle (UAV) operators exhibit far lower rates of post-traumatic stress disorder than soldiers who are engaged in close-up combat.[47] Indeed, scholarship is replete with anecdotes about "cubicle warriors"[48] exhibiting a "PlayStation mentality" in which they either give no thought to the implications of their killing or, even worse, actually enjoy it. For example, one UAV pilot told Peter Singer that remote combat is "like a video game. It can get a little bloodthirsty. But it's fucking cool," while another admitted that although he "thought killing somebody would be this life-changing experience," he actually found "[k]illing people is like squashing an ant. I mean, you kill somebody and it's like 'All right, let's go get some pizza.'"[49]

The burdened-psyche objection also ignores the humans who operate autonomous weapons.[50] As we have seen, an AWS carries out the intent of its operators even when attacking a general class of targets instead of a specific target or group of targets. There is no reason to believe that the human who activates or programs an autonomous weapon does not understand the gravity of his decision to activate the machine or does not have a burdened psyche when the machine does what is expected of it—kill.[51] To be sure, that does not mean an AWS programmer or activator will have the same emotional reaction to killing as a soldier who kills in close combat or even the same emotional reaction as a UAV operator who can at least see the person he kills on his video screen. But that is a difference of degree, not of kind—and it is entirely possible that an AWS programmer or activator will

complete the mission efficiently and with the least possible amount of casualties"); LIM, *supra* note 45, at 174 ("Instead of killing as the result of an emotional process that recognizes the gravity of the situation, human combatants may be emotionally overwhelmed by an instinct to survive. Surely, killings in service of self-preservation can hardly be considered respectful of the victim.").

[47] *See, e.g.*, Wayne Chappelle et al., *Combat and Operational Risk Factors for Post-Traumatic Stress Disorder Symptom Criteria Among United States Air Force Remotely Piloted Aircraft "Drone" Warfighters*, 62 J. ANXIETY DISORDERS 86, 91 (2019). In addition, one study has found that pilots of manned aircraft suffer even fewer psychiatric symptoms than UAV operators. *See generally* Rajiv Kuman Saini et al., *Cry in the Sky: Psychological Impact on Drone Operators*, 30 IND. PSYCH. J. S15–S19 (2021). *But see* Jean L. Otto & Bryant J. Webber, *Mental Health Diagnoses and Counseling among Pilots of Remotely Piloted Aircraft in the United States Air Force*, 20 MED. SURVEILLANCE MONTHLY REP. 3, 3 (2013) (finding no relevant differences).

[48] Lambèr Royakkers & Rinie van Est, *The Cubicle Warrior: The Marionette of Digitalized Warfare*, 12 ETHICS INF. TECH. 289, 290 (2010).

[49] *Id.*

[50] Robert Sparrow, *Robotic Weapons and the Future of War*, *in* NEW WARS AND NEW SOLDIERS: MILITARY ETHICS IN THE CONTEMPORARY WORLD 117, 125 (Paolo Tripodi & Jessica Wolfendale eds., 2012) (noting that "it might be argued that the proper place to look for the required attitude is rather in the person who ordered the deployment of the AWS").

[51] *Cf.* Michael C. Horowitz, *The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons*, 145 DAEDALUS 25, 31 (2016) ("By ensuring that potential operators of LAWS understand how they operate—and feel personally accountable for their use—militaries can theoretically avoid offloading moral responsibility for the use of force.").

feel a particular psychological burden knowing that he is responsible for the actions of an uncommonly powerful weapon. Regardless, the involvement of the programmer or activator indicates that there *is* at least one human who can suffer from the "decision" of an autonomous weapon to take human life.[52]

C. Only Humans Risk

The third deontological objection to autonomous weapons is that killing in combat is ethical only if the combatant risks his own life. Some scholars phrase this objection as a matter of military honor. Aaron Johnson and Sidney Axinn, for example, claim that the creation of honor "requires that humans risk sacrifice. Where there is no human in the loop, there is no one to risk sacrifice, and therefore no honor produced."[53] Others phrase the objection in terms of the relationship between soldier and target. Ozlem Ulgen, for example, argues that "[b]y replacing the human combatant with a machine the combatant's human dignity is not only preserved but elevated above the human target. This can also be seen as a relative end in that it selfishly protects your own combatants from harm at all costs."[54]

Both of these quotes conceptualize "ethical" combat as a physical affair in which combatants directly encounter each other. Indeed, Ulgen openly states that "without face-to-face killing certain humans are deemed more valuable and priceless than others, which creates a hierarchy of human dignity."[55] This is, however, an anachronistic understanding of combat. Although nearly all armed conflicts continue to involve some amount of face-to-face fighting, "[i]n the twenty-first century, remote warfare has been the most common form of military engagement used by states."[56] The entire point of remote warfare—high-altitude bombing, over-the-horizon weapons, UAVs, even snipers—is to protect combatants from harm while they are harming the enemy. So, if it is unethical to kill unless a soldier risks sacrificing his own life, nearly all killings in combat today are just as unethical as killings carried out by autonomous weapons.[57] The objection

[52] Sparrow, *supra* note 50, at 125 ("If there can be an interpersonal relationship between a bombardier and the enemy combatants they target thousands of feet below, it is not clear why a similar relationship could not exist between the person who orders the deployment of an autonomous weapon system and the people who are killed by that system.").

[53] Aaron M. Johnson & Sidney Axinn, *The Morality of Autonomous Robots*, 12 J. MIL. ETHICS 129, 136 (2013).

[54] Ozlem Ulgen, *Human Dignity in an Age of Autonomous Weapons: Are We in Danger of Losing an 'Elementary Consideration of Humanity'?*, 17 BALTIC Y.B. INT'L L. ONLINE 167, 177 (2020).

[55] *Id.* at 175.

[56] Abigail Watson & Alasdair McKay, *Remote Warfare: A Critical Introduction*, E-INT'L REL. 1 (2021), https://www.e-ir.info/2021/02/11/remote-warfare-a-critical-introduction/ [https://perma.cc/8T6F-FMGL].

[57] *See, e.g.*, Lieblich & Benvenisti, *supra* note 22, at 257 ("[T]his objection does not capture the salient dilemma of AWS because it could equally apply also to other methods and tactics of warfare such as drones, cruise missiles and high-altitude bombing.").

about sacrifice, therefore, is less an objection to AWS than to modern warfare itself.[58]

Moreover, the basic idea underlying the objection—that a soldier cannot ethically kill unless he risks his own life while doing so—is anything but self-evident. Why is it "selfish" for a state to protect its soldiers from harm, as Ulgen argues? The alternative, requiring a state to expose its soldiers to avoidable risks because the enemy lacks its technological sophistication, simply leads to more unnecessary death on the battlefield—death suffered by ordinary soldiers, many of whom will not have chosen to fight in the first place. Demanding they die to preserve some medieval notion of chivalric combat hardly seems ethical—and explicitly runs afoul of at least some just-war thinking, such as Bradley Jay Strawser's "principle of unnecessary risk," according to which "it is wrong to command someone to take on unnecessary potentially lethal risks in an effort to carry out a just action for some good."[59] Indeed, Strawser has persuasively argued that the principle of unnecessary risk imposes an ethical obligation to engage in remote warfare whenever possible.[60]

Johnson and Axinn justify the need for soldiers to risk dying in combat—the necessary condition of acting with honor—somewhat differently than Ulgen. In their view, "honorable behavior is a useful war strategy, as well as a moral requirement," because "[i]f a nation behaves dishonorably, by ignoring the laws of warfare or simple humanitarian matters, their enemy may hate them so much that peace cannot be arranged for a very long time."[61] That is no doubt true, which is why a number of militaries specifically define honor as the willingness and ability to conduct hostilities in a manner that complies with IHL.[62] Neither complying with IHL nor acting in a generally humanitarian manner, however, requires soldiers to be prepared to sacrifice their lives in combat. So if autonomous weapons ever become capable of achieving both of those goals as well as human soldiers, Johnson and Axinn's sacrifice argument will be nothing more than a demand that militaries needlessly squander human lives.

D. Only Humans Can Kill with Dignity

The fourth and final deontological objection to autonomous weapons is that machine killing is inherently unethical because it dehumanizes the victim. This

[58] *See, e.g.*, Dieter Birnbacher, *Are Autonomous Weapons Systems a Threat to Human Dignity?*, *in* AUTONOMOUS WEAPONS SYSTEMS: LAW, ETHICS, POLICY 105, 119 (Nehal Bhuta et al. eds., 2016) (noting that "all of the features that make AWS appear to be problematic from the viewpoint of human dignity might be present in conventional acts of war").

[59] Bradley Jay Strawser, *Moral Predators: The Duty to Employ Uninhabited Aerial Vehicles*, 9 J. MIL. ETHICS 342, 344 (2010).

[60] *Id*. at 343.

[61] Johnson & Axinn, *supra* note 53, at 133, 136.

[62] DEPARTMENT OF DEFENSE, LAW OF WAR MANUAL 67 (2015) (stating that honor requires that "parties to a conflict must accept that certain limits exist on their ability to conduct hostilities" such as that "the right of belligerents to adopt means of injuring the enemy is not unlimited").

objection takes two forms, one perpetrator-centered and the other victim-centered. The perpetrator-centered version, echoing the moral-judgment objection, focuses on the inability of an AWS to recognize the humanity of the individual it kills. Thus, Rosert and Sauer claim that "[t]reating a human as an object is what happens when LAWS are allowed to kill. The victim, be she combatant or civilian, is reduced to a data point in an automated killing machinery that has no conception of what it means to take a human life,"[63] while Heyns insists that "[t]o allow such machines to determine whether force is to be deployed against a human being may be tantamount to treating that particular individual not as a human being but, rather, as an object eligible for mechanized targeting."[64] The victim-centered version, by contrast, focuses on the subjective experience of the individuals targeted by machines. Citing data about UAV killings in Yemen, Ulgen argues that human targets care about "whether they are killed by autonomous weapons or soldiers," suffering greater psychological harm when it is the former.[65]

The problem with the perpetrator-centered version of this objection is that, like the previous objections, it does not apply only to autonomous weapons. On the contrary, "mechanized targeting" in which the victim is "reduced to a data point in an automated killing machinery" is the defining characteristic of most modern forms of remote warfare.[66] The individuals killed by bombs dropped by a B-52 bomber at 25,000 feet are no less nameless, faceless objects than the individuals killed by an AWS. The same is true of combatants killed by a Brimstone "fire and forget" missile seeking out a tank with a particular millimeter-wave signature,[67] by artillery conducting radar-directed counterbattery fire,[68] or by infantry engaging in long-range fire or barrage fire. It is even true of most of the individuals who are

[63] Rosert & Sauer, *supra* note 19, at 372; *see also* Problems with Autonomous Weapons, *supra* note **Error! Bookmark not defined.** ("Killer robots . . . challenge human control over the use of force, and where they target people, they dehumanise us—reducing us to data points.").

[64] Heyns, *supra* note 17, at 10; *see also* Johnson & Axinn, *supra* note 53, at 134 ("A mouse can be caught in a mousetrap, but a human must be treated with more dignity . . . A robot is in a way like a high-tech mousetrap; it is not a soldier with concerns about human dignity or military honor. Therefore, a human should not be killed by a machine as it would be a violation of our inherent dignity.").

[65] Ulgen, *supra* note 54, at 183; *cf.* Jens David Ohlin, *Autonomous Weapons and Reactive Attitudes*, *in* LETHAL AUTONOMOUS WEAPONS: RE-EXAMINING THE LAW AND ETHICS OF ROBOTIC WARFARE 189, 193 (Jai Galliott, Duncan MacIntosh, & Jens David Ohlin eds., 2021) ("[I]t will be extremely hard for the victim to resist certain reactive attitudes, including feelings of resentment, as long as the AWS is sophisticated enough that its behavior seems similar to the decisions that a reasonably law-compliant human agent would make.").

[66] Rosert & Sauer, *supra* note 19, at 372.

[67] David Hambling, *New British Brimstone 2 Missiles Are Bad News For Russian Tanks, Artillery, Air Defense And Command Posts*, FORBES (Nov. 28, 2022), https://www.forbes.com/sites/davidhambling/2022/11/28/new-british-brimstone-2-missiles-are-bad-news-for-russian-tanks-artillery-air-defense-command-posts/ [https://perma.cc/WL5F-58B4].

[68] *See* BAKER, *supra* note 22, at 43 (noting that "[s]uch a counterbarrage will seek to hit not only the artillery pieces the enemy is using to fire the incoming artillery shells, but also any enemy personnel, vehicles and equipment in the unit that operates those artillery pieces, even though the commander of the battery in question may well not know the specifics of those additional targets").

killed by UAVs, given that the vast majority of drone killings are "signature strikes"[69] targeting "groups of men who bear certain signatures, or defining characteristics associated with terrorist activity, but whose identities aren't known."[70] To be sure, the soldiers responsible for delivering such remote death may be aware that their actions will have lethal consequences. But their awareness does not seem any more immediate or less objectifying than the awareness of the commander who activates an autonomous weapon.

The victim-centered version of the "dehumanization" objection to autonomous weapons fares little better. The UAV statistics Ulgen cites indicate that humans fear being killed by machines more than by their fellow humans—not that humans fear being killed by autonomous machines more than by non-autonomous ones. In terms of long-distance killing, the more specific claim cannot be true: a victim would have no way of knowing whether the bomb, missile, or artillery shell heralding her imminent death was fired by a human or by a machine. At best, then, the victim-centered objection makes sense only when AWS directly participate in face-to-face combat.

But even in that context the objection is unpersuasive. To begin with, the recognition problem still applies, because "autonomous and remotely piloted systems will probably appear alongside each other on future battlefields and . . . many systems will be designed so as to be able to be switched from remotely piloted mode to autonomous mode."[71] The objection thus imagines autonomous weapons of the Terminator variety—scary-looking robots moving and killing of their own accord—which do not currently exist and likely never will.

Even if victims could tell the difference between manned and unmanned weapons systems on the battlefield, it is still not self-evident they would fear unmanned weapons more. Most obviously, given that all humans are driven by self-preservation, many are likely indifferent to whether they are killed by another human or by a machine. "[S]eeing the man's eyes as he stabs you doesn't make your death any more palatable."[72] Moreover, even if some people are not indifferent to the manner of their death, there are a number of reasons why they might prefer to be killed by a machine instead of by another human. Duncan MacIntosh offers one possibility, which inverts a critique of autonomous weapons that is discussed

[69] *See* Kevin Jon Heller, *"One Hell of a Killing Machine": Signature Strikes and International Law*, 11 J. INT'L CRIM. JUST. 89, 90 (2013).
[70] DANIEL KLAIDMAN, KILL OR CAPTURE: THE WAR ON TERROR AND THE SOUL OF THE OBAMA PRESIDENCY 41 (2012).
[71] BAKER, *supra* note 22, at 107–08. An example is the Optionally Manned Fighting Vehicle, designed and manufactured by General Dynamics, which is scheduled to replace the Bradley Fighting Vehicle in 2030. *See Army Announces Contract Awards for OMFV*, ARMY PUBLIC AFFAIRS (June 26, 2023), https://www.army.mil/article/267920/army_announces_contract_awards_for_omfv [https://perma.cc/BSQ2-JERT].
[72] Gregory P. Noone & Diana C. Noone, *The Debate Over Autonomous Weapons Systems*, 47 CASE WESTERN RES. J. INT'L L. 25, 33 (2015).

in Section III: "if it is a human who is killing you, you might experience not only the horror of your pending death, but also anguish at the fact that, even though they could take pity on you and spare you, they will not—they are immune to your pleading and suffering."[73] Jens David Ohlin provides another: "the AWS would do its work dispassionately and without illegitimate motivations such as discrimination on the basis of race, religion, or nationality."[74] In both situations the non-human nature of autonomous weapons would be a virtue, not a vice.

There is, however, an even more compelling reason why a victim might prefer to be killed by a machine than by a human, one that also reflects autonomous weapons' lack of humanity: their superior accuracy. The victim-centered dehumanization objection presumes that victims will both expect and receive a quick, clean death at the hands of a human soldier. But such deaths seem more the exception than the rule: there is nothing quick or clean about being burned alive by a flamethrower, stabbed to death by a bayonet, or slowly bleeding out from a fatal but poorly aimed bullet. That prospect is surely more terrifying than the likelihood—due to advanced technology—of "being shot through the head or heart and instantly killed by a machine."[75]

E. The Limits of Deontology

In short, none of the deontological objections to autonomous weapons are persuasive. All four share a common weakness: they apply not only to AWS, but also to a wide variety of weapons that states have been using for decades without significant ethical objection. Moreover, the one relevant difference between autonomous weapons and weapons like high-altitude bombers and drones—the former's greater accuracy—counts in favor of killer robots, not against them.

The most fundamental problem with deontological objections, however, is precisely their deontology. Because deontological arguments are by definition non-consequentialist, they would prohibit states from using autonomous weapons in war *even if doing so would lead to fewer civilian casualties and less unnecessary combatant suffering*. Indeed, most deontologists openly assert that it is morally irrelevant whether AWS could comply with IHL as well as—or even better than—human soldiers. Alex Leveringhaus claims, for example, that "[o]ne could imagine a technologically perfect Killer Robot, potentially capable of fully complying with the three principles of *jus in bello*, yet oppose this type of weapon."[76] Similarly, Amoroso et al. insist that "respect for human dignity affords a distinctive moral reason to forbid the use of AWS, which cannot be overridden by any envisaged

[73] Duncan MacIntosh, *Fire and Forget: A Moral Defense of the Use of Autonomous Weapons Systems in War and Peace*, *in* LETHAL AUTONOMOUS WEAPONS: RE-EXAMINING THE LAW AND ETHICS OF ROBOTIC WARFARE 9, 19 (Jai Galliott, Duncan MacIntosh & Jens David Ohlin eds., 2021).

[74] Ohlin, *supra* note 65, at 191–92.

[75] Horowitz, *supra* note 51, at 32.

[76] Alex Leveringhaus, *What's So Bad About Killer Robots?*, 35 J. APPL. PHILOS. 341, 343 (2018).

technological developments that may occur in the future, even by technological developments that might lead to improved performances in AWS's critical targeting and engagement functions."[77] And Robin Geiss is perhaps the most uncompromising of all, suggesting that "[t]he inherent irrationality that is always part and parcel of a human decision to kill could itself be regarded as a prerequisite for at least a minimum degree of moral substance."[78]

There is no way to rebut such arguments. One either accepts or rejects the premises they are based on, such as that being killed by a machine is dehumanizing and that avoiding dehumanization is a greater good than not being erroneously killed. Nevertheless, most people who are leery of autonomous weapons would likely still agree with more pragmatic deontologists like Sparrow, who acknowledge that "[i]f AWS would kill fewer noncombatants than human troops, this establishes a strong consequentialist case for their deployment, regardless of other ethical concerns about them."[79] Indeed, as MacIntosh says, if "given a choice between control by a morally bad human who would kill someone undeserving of being killed and a morally good robot who would kill only someone deserving of being killed," most people "would pick the good robot."[80]

Perhaps recognizing the limits of such uncompromising deontology, most critics of autonomous weapons object to them on consequentialist grounds. In their view, regardless of how precise AWS may be in terms of targeting, permitting states to use them will make war *less* humane, not *more*. Heyns expresses this argument well:

> Yet robots have limitations in other respects as compared to humans. Armed conflict and IHL often require human judgement, common sense, appreciation of the larger picture, understanding of the intentions behind people's actions, and understanding of values and anticipation of the direction in which events are unfolding. Decisions over life and death in armed conflict may require compassion and intuition. Humans—while they are fallible—at least might possess these qualities, whereas robots definitely do not.[81]

[77] AMOROSO ET AL., *supra* note 18, at 33.

[78] ROBIN GEISS, THE INTERNATIONAL-LAW DIMENSION OF AUTONOMOUS WEAPONS SYSTEMS 17 (2015); *see also* Skerker, Purves, & Jenkins, *supra* note 14, at 207–08 (insisting that even a "sophisticated AWS" that "might adhere at least as well as human combatants to the principles of just war because it makes fewer empirical and practical mistakes" would lack the moral agency necessary to use lethal force).

[79] Sparrow, *supra* note 16, at 102; *see also* Purves, Jenkins & Strawser, *supra* note 20, at 867 ("[I]f deploying AWS in a particular conflict can be expected to reduce civilian casualties from 10,000 to 1000, this consideration might very well override the fact that AWS would not act for the right reasons in achieving this morally desirable result.").

[80] MacIntosh, *supra* note 73, at 10.

[81] Heyns, *Report, supra* note 1, at ¶ 55.

Heyns is on strong ground when he claims that autonomous weapons currently lack these quintessentially human qualities. Moreover, although the idea that AWS will eventually possess them cannot be ruled out, programming a machine to demonstrate (say) common sense would likely require artificial general intelligence, which remains decades away—if it is ever achieved.[82]

By itself, however, the inhumanity of autonomous weapons does not support the consequentialist objection. Critics also need to show that, because they lack human qualities like "common sense" and "compassion," AWS will never be able to comply with IHL as well as human soldiers. That is the relevant standard, as even most consequentialist critics of AWS,[83] including Heyns,[84] acknowledge. If autonomous weapons ever comply with IHL at least as well as humans, the only possible objection to their use in armed conflict would be deontological.

As the next section of the article explains, there is every reason to believe that autonomous weapons will eventually be able to satisfy this consequentialist standard, making their use ethically permissible—and legally required.[85] Part of the explanation focuses on the machine side of the equation: the technological ability of AWS to recognize objects and target with precision will only increase over time. But the human side of the equation is even more important, because the consequentialist objection to AWS ignores the basic lesson of decades of research in cognitive psychology: humans are extremely bad at making the kind of rational judgments that complying with IHL requires, particularly when they find themselves in dangerous and uncertain situations like combat.

[82] *See, e.g.*, BOULANIN & VERBRUGGEN, *supra* note 3, at 92 (noting that artificial general intelligence "does not currently exist and remains for now in the realm of science fiction").

[83] *See, e.g.*, Amoroso & Tamburrini, *supra* note 15, at 5–6 ("An AWS complying with IHL requirements is usually taken to be an autonomous weapon which is capable of respecting the principles of distinction and proportionality at least as well as a competent and conscientious human soldier."); Robert Sparrow, *Twenty Seconds to Comply: Autonomous Weapon Systems and the Recognition of Surrender*, 91 INT'L L. STUD. 699, 711 (2015) ("To insist on the reasonable expectation standard is just to insist that we do not owe surrendered combatants any less when we send a robot rather than a human being into combat.").

[84] *See* Heyns, *Report*, *supra* note 1, ¶ 65 ("A consideration of a different kind is that if it is technically possible to programme LARs to comply better with IHL than the human alternatives, there could in fact be an obligation to use them.").

[85] Art. 57(1) of the First Additional Protocol provides that "[i]n the conduct of military operations, constant care . . . be taken in the conduct of military operations to spare the civilian population." Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), rt. 57(1), *adopted* June 8, 1977, 1125 U.N.T.S. 3 [hereinafter AP I]. If a state had the ability to minimise civilian casualties by using an autonomous weapon instead of human soldiers, that provision would require the state to use the AWS.

II. The Necessity of Human Understanding

Critics of autonomous weapons normally focus on two of the most basic principles of IHL, distinction and proportionality.[86]

A. Distinction

The principle of distinction, codified in Article 48 of the First Additional Protocol (AP I),[87] prohibits intentionally attacking civilians. Distinction-based arguments against autonomous weapons invariably focus on the inability of machines to determine the intentions of the individuals they target. Human Rights Watch, for example, says that "fully autonomous weapons would not possess human qualities necessary to assess an individual's intentions, an assessment that is key to distinguishing targets," because "[o]ne way to determine intention is to understand an individual's emotional state, something that can only be done if the soldier has emotions."[88] Similarly, Marcello Guarini and Paul Bello insist that "[a] robot without representation of or the ability to recognize these emotional states would be at a crippling disadvantage in the battlefield, especially if its task requires dealing with noncombatants or others whose status has to be determined," because "a robot that cannot tell the difference between fear and anger will have a very hard time assessing the intent of an agent."[89]

The idea that the principle of distinction normally requires an attacker to discern a target's intent is significantly overstated. In many situations, targetability is determined solely on the basis of objective, externally manifested signs and behavior, rendering what is going on in the target's mind irrelevant.[90] The most obvious such situation involves combatants who distinguish themselves from the civilian population in an international armed conflict by wearing a uniform (members of a state's regular armed forces) or a fixed and distinctive sign recognizable at a distance ("members of other militias and members of other volunteer corps, including those of organized resistance movements"), the basic

[86] *See* Noel Sharkey, *Saying 'No!' to Lethal Autonomous Targeting*, 9 J. MIL. ETHICS 369, 378 (2010).

[87] AP I, *supra* note 85, Art. 48 ("In order to ensure respect for and protection of the civilian population and civilian objects, the Parties to the conflict shall at all times distinguish between the civilian population and combatants and between civilian objects and military objectives and accordingly shall direct their operations only against military objectives.").

[88] HUMAN RIGHTS WATCH, LOSING HUMANITY: THE CASE AGAINST KILLER ROBOTS 31 (2012).

[89] Marcello Guarini & Paul Bello, *Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 129, 137–38 (Patrick Lin, Keith Abney & George A. Bekey eds., 2012).

[90] *Cf.* Marco Sassóli, *Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified*, 90 INT'L L. STUD. 308, 333 (2014) ("Even a human being engaged in hostilities will never know, and is not required to inquire into, the intent of another human being, but instead will be receptive only to objective indications of the danger a person represents.").

requirement for POW status upon capture.[91] Lawfully targeting combatants in those categories requires nothing more than the ability to identify clothing as a uniform or a marking as a "fixed distinctive sign."[92] Such identification is an object-recognition task; no intent assessment is necessary. Indeed, if an individual is wearing the uniform of a party to the conflict or the fixed and distinctive sign of a militia, IHL permits him to be targeted anywhere, anytime, with any amount of force.[93]

This is an important point, because there is no reason to believe human soldiers will always be better than autonomous weapons at recognizing uniforms or fixed and distinctive signs. As Elliot Winter notes, not only can machines "observe at least as well as humans and, indeed, at higher resolution and with greater rapidity,"[94] their recognition ability "has now advanced to a point where it has reached parity with human recognition abilities."[95] For example, Malong Technologies, a Chinese company, has developed an AI-based system that can classify millions of photos of consumer objects, including clothing, into a thousand categories with 94.78% accuracy—essentially the same as human performance.[96]

Because critics generally acknowledge the possibility that autonomous weapons will be able to comply with the principle of distinction when targets are lawful combatants,[97] they tend to focus on situations that are ostensibly more complex. In particular, a number of critics have argued that AWS will never be able

[91] Geneva Convention (III) Relative to the Treatment of Prisoners of War, art. 4(A), Aug. 12, 1949, 75 U.N.T.S. 135.

[92] *Id.*

[93] *See, e.g.*, GARY D. SOLIS, THE LAW OF ARMED CONFLICT INTERNATIONAL HUMANITARIAN LAW IN WAR 188 (2nd ed. 2018) ("When a soldier is bivouacked and sleeping she remains a combatant and so remains a legitimate target. While sleeping, she may be lawfully killed by an opposing lawful combatant. If a combatant is targeted far behind the front lines, no matter how unlikely such targeting may be, she continues to be a legitimate target for opposing lawful combatants."); Geoffrey S. Corn, *Autonomous Weapons Systems: Managing the Inevitability of "Taking the Man out of the Loop"*, *in* AUTONOMOUS WEAPONS SYSTEMS: LAW, ETHICS, POLICY 209, 230 (Nehal Bhuta et al. eds., 2016) ("[U]se-of-force authority is based on the presumptive threat posed by members of the enemy group and not on individualized conduct-based threat validation. Since such members represent a presumed threat unless and until rendered *hors de combat*, attacking forces are legally justified in employing deadly combat power against such members as a measure of first resort.").

[94] Elliot Winter, *The Compatibility of Autonomous Weapons with the Principle of Distinction in the Law of Armed Conflict*, 69 ICLQ 845, 859 (2020).

[95] *Id.* at 867; *see also* CENTER FOR A NEW AMERICAN SECURITY, AUTONOMOUS WEAPONS AND HUMAN CONTROL 5 (2016) ("It is true that machines' abilities at object recognition are rapidly improving and may soon surpass humans' ability to accurately identify objects.").

[96] Winter, *supra* note 94, at 867.

[97] *See, e.g.*, Christof Heyns, *Autonomous Weapons in Armed Conflict and the Right to a Dignified Life: An African Perspective*, 33 SOUTH AFRICAN J. ON HUM. RTS. 46, 53 (2017) ("Autonomous weapons may find it easier, specifically, to identify status-based targets, such as members of a declared hostile force, as is often the case in international armed conflicts."); HUMAN RIGHTS WATCH, *supra* note 88, at 30 (acknowledging that recognition problems will be less significant when combatants are wearing "uniforms or insignia").

to reliably recognize when a combatant is surrendering—a situation that can occur in any warfighting domain[98]—because such recognition requires the human ability to determine that the combatant is intending to lay down his arms.[99] According to Sparrow, for example, "[t]he actions that indicate surrender vary with context, both internationally, and also amongst different types of military units. For this reason—and given the possibility that the forces involved in a conflict may be operating with different understandings as to the relevant conventions—recognizing surrender is fundamentally a question of recognizing an intention."[100]

This is not correct. A combatant does not have to subjectively intend to surrender to be properly deemed *hors de combat* under Art. 41(1) of AP I.[101] On the contrary, as the ICRC Commentary to Art. 41 makes clear, a combatant becomes *hors de combat* only if he "clearly expresses an intention to surrender" by engaging in one of a limited number of internationally recognized behaviors:

> In general, a soldier who wishes to indicate that he is no longer capable of engaging in combat, or that he intends to cease combat, lays down his arms and raises his hands. Another way is to cease fire, wave a white flag and emerge from a shelter with hands raised . . . In the air, it is generally accepted that a crew wishing to indicate their intention to cease combat, should do so by waggling the wings while opening the cockpit (if this is possible). At sea, fire should cease and the flag should be lowered.[102]

In other words, for purposes of surrender, what matters is *how the combatant acts*, not *what the combatant thinks*.[103] This distinction is fundamental: if a combatant does not externally manifest his intention to surrender by acting in one of the ways mentioned in the ICRC Commentary, he remains a lawful target even if, in his heart of hearts, he genuinely intends to stop fighting. Similarly, if the

[98] AMOROSO ET AL., *supra* note 18, at 25 (noting, with regard to surrender, that "this rule applies to every warfare scenario in which humans are involved. It therefore counters the argument that IHL would not pose any obstacle to the deployment of lethal AWS in a variety of scenarios where civilians or civilian objects are totally absent (e.g. a battleship in the high seas)").

[99] *See, e.g.*, *id.* ("[T]he recognition of behaviors that convey unconventional surrender messages and fighting incapacitation poses formidable challenges for AWS programmers and developers."); Heyns, *supra* note 1, ¶ 67 ("It would be difficult for robots to establish, for example . . . whether soldiers are in the process of surrendering."); *cf.* HUMAN RIGHTS WATCH, *supra* note 88, at 34 ("Identifying whether an enemy soldier has become *hors de combat*, for example, demands human judgment.")

[100] Sparrow, *supra* note 83, at 707.

[101] AP I, *supra* note 85, Art. 41(1) ("A person who is recognized or who, in the circumstances, should be recognized to be *hors de combat* shall not be made the object of attack. (2) A person is *hors de combat* if . . . (b) he clearly expresses an intention to surrender.").

[102] YVES SANDOZ ET AL. (EDS.), COMMENTARY ON THE ADDITIONAL PROTOCOLS OF 8 JUNE 1977 TO THE GENEVA CONVENTIONS OF 12 AUGUST 1949, 486–87 (1987).

[103] *See* Sassóli, *supra* note 90, at 315 ("What counts, for example, is, not whether a person wants to surrender, but whether he or she indicates their willingness to surrender and the attacker becomes aware of this indication.").

combatant does "clearly express his intention to surrender" by acting in one of the recognized ways, he cannot be attacked even if enemy soldiers suspect that he will start shooting again once they attempt to capture him—an act of perfidy.[104] They must treat him as *hors de combat*, despite their suspicions, until his actions make clear that he does not actually intend to surrender, such as by picking up his weapon again.[105]

Because *hors de combat* status depends on how a combatant acts, not what a combatant thinks, there is no reason to believe that autonomous weapons will never be able to reliably determine whether a combatant is surrendering. That determination is an object-recognition task that involves a limited number of behaviors (such as raising hands or waving a white flag), much like the object-recognition task involved in determining whether an individual is a combatant (wearing a uniform or a fixed and distinctive sign)—a task clearly within the capabilities of autonomous weapons,[106] as even their critics acknowledge.

Perhaps aware of the surrender argument's limitations, some critics of autonomous weapons emphasize the scenario mentioned above, where a clearly expressed intention to surrender is perfidious. Sparrow, for example, claims that the possibility of perfidy means machines "must be capable of distinguishing between real and feigned intentions," a task "significantly more difficult than the task of recognizing the signal in the first place."[107] That is no doubt true, and the recognition problem Sparrow identifies is not limited to surrender. All forms of perfidy—such as feigning incapacitation by wounds or feigning civilian status by not wearing a uniform—require soldiers to understand their enemy's "true" intentions. The question, though, is not whether autonomous weapons can reliably determine when a combatant waving a white flag actually intends to attack or when an individual wearing civilian clothes is actually a combatant. The question is whether AWS cannot make those determinations *as well as human soldiers*. If humans struggle to distinguish "between real and feigned intentions" just as much as machines, the evident possibility of perfidy is not a viable argument against autonomous weapons.

There is no obvious way to compare the ability of humans and machines to recognize perfidy, but there is also no reason to believe significant differences exist between them. First, as Michael N. Schmitt notes, "asymmetrically disadvantaged enemies have been feigning civilian or other protected status to avoid being

[104] Perfidy is defined as an act "inviting the confidence of an adversary to lead him to believe that he is entitled to, or obliged to accord, protection under the rules of international law applicable in armed conflict, with intent to betray that confidence." AP I, *supra* note 85, art. 37(1).

[105] *Hors de combat* status continues to exist only insofar as the combatant "abstains from any hostile act." *Id*., art. 41.

[106] *See* Nathan Gabriel Wood, *Autonomous Weapon Systems and Responsibility Gaps: A Taxonomy*, 25 ETHICS & INFO. TECH. 1, 7 n.22 (2023) (noting that "currently existing AWS are fully capable of recognizing hands held high as an indication of surrender").

[107] Sparrow, *supra* note 83, at 708.

engaged by human-operated weapon systems for decades (even centuries)."[108] If perfidy did not work against humans—if they were generally able to distinguish between real and feigned intentions—perfidy would have died out long ago, because it generally requires a combatant to expose himself to attack, such as by setting down his weapon. Second, given that perfidious combatants are unlikely to communicate their true intent, soldiers must infer that intent from the combatant's actions. A reliable indicator of feigned surrender, for example, is an enemy soldier reaching for the weapon he put down as soon as the enemy comes into firing range. There is no reason to believe an AWS could not recognize and classify the act of reaching for a weapon as quickly and accurately as a human soldier—and lacking an instinct for self-preservation, an autonomous weapon would almost certainly wait longer, acquiring more relevant information, before opening fire.

Another distinction situation often mentioned by critics of autonomous weapons poses a greater challenge: direct participation in hostilities (DPH). A civilian who directly participates in hostilities loses protection from attack for the duration of his direct participation.[109] Lawfully targeting this category of individuals thus requires an attacker to be able to determine whether the target's actions qualify as direct participation—defined by the ICRC as "specific hostile acts carried out by individuals as part of the conduct of hostilities between parties to an armed conflict."[110]

According to the critics, this test requires the uniquely human ability to discern an individual's intention. Sharkey, for example, says that "[i]n a war with non-uniformed combatants, knowing who to kill would have to be based on situational awareness and on having human understanding of other people's intentions and their likely behaviour. In other words, human inference is required. Humans understand one another in a way that machines cannot."[111] Similarly, Guarini and Bello claim that "[w]ithout reliable intentional state attribution, it is hard to see how a robot could usefully assess threatening from nonthreatening behavior, and without that, distinguishing combatants from noncombatants will be exceedingly difficult."[112]

[108] Michael N. Schmitt, *Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics*, HARV. NAT'L SEC. J. FEATURES 1, 12 (2013).

[109] ICRC, *Direct Participation in Hostilities: Questions & Answers*, https://www.icrc.org/en/doc/resources/documents/faq/direct-participation-ihl-faq-020609.htm [https://perma.cc/6STJ-CQWT].

[110] INTERNATIONAL COMMITTEE OF THE RED CROSS, INTERPRETIVE GUIDANCE ON THE NOTION OF DIRECT PARTICIPATION IN HOSTILITIES UNDER INTERNATIONAL HUMANITARIAN LAW 45 (2009) [hereinafter ICRC INTERPRETIVE GUIDANCE].

[111] Sharkey, *supra* note 86, at 379.

[112] Guarini & Bello, *supra* note 89, at 134; *cf.* Markus Wagner, *The Dehumanization of International Humanitarian Law: Legal, Ethical, and Political Implications of Autonomous Weapon Systems*, 47 VAND. TRANS. J. INT'L L. 1371, 1392 ("Not only would AWS have to be able to distinguish civilians from military personnel, but it must also decide if a civilian is taking a 'direct part in hostilities.' These situations are challenging for humans to judge, and it does not appear that the necessary contextual analysis is amenable to easily programmable quantitative assessments at this time."); Heyns, *supra* note 1, ¶ 68 ("Experts have noted that for counter-

Once again, the emphasis on subjective intent is overstated. In many situations, it will be possible to determine whether a civilian is directly participating in hostilities solely on the basis of his "specific hostile acts";[113] the intention behind those acts will be irrelevant.[114] That is true of a number of examples of direct participation provided by the ICRC, such as firing at the enemy, identifying and marking targets, and delivering ammunition to soldiers on the front line.[115] A civilian who engages in any of these acts is targetable regardless of the intent behind the act, because the objective qualities of the act itself satisfy the DPH test.[116]

Indeed, the ICRC goes to great lengths to make clear that direct participation does not depend on subjective intent. Unlike threshold of harm and direct causation, which are clearly objective requirements, belligerent nexus could be interpreted subjectively, given that it requires the specific hostile act "be specifically designed" to support one of the parties to the conflict.[117] The ICRC, however, specifically disavows any such interpretation:

> Belligerent nexus should be distinguished from concepts such as subjective and hostile intent. These relate to the state of mind of the person concerned, whereas belligerent nexus relates to the objective purpose of the act. That purpose is expressed in the design of the act or operation and does not depend on the mindset of every participating individual. As an objective criterion linked to the act alone, belligerent nexus is generally not influenced by factors such as personal distress or preferences, or by the mental ability or willingness of persons to assume responsibility for their conduct.[118]

The word "generally" is important here. Although most acts of direct participation do not require knowledge of the civilian's intention, in some situations intent remains relevant. Human Rights Watch offers one hypothetical example: a frightened mother runs after her two children, yelling at them to stop playing with toy guns near a soldier. According to the organization, an autonomous weapon would be far more likely than a human soldier to mistakenly attack the children, because "[a] human soldier could identify with the mother's fear and the children's

insurgency and unconventional warfare, in which combatants are often only identifiable through the interpretation of conduct, the inability of LARs to interpret intentions and emotions will be a significant obstacle to compliance with the rule of distinction.").

[113] ICRC INTERPRETIVE GUIDANCE, *supra* note 110, at 43.

[114] *Cf.* Sassóli, *supra* note 90, at 315 ("It would be a misconception of existing IHL to claim that the decision to kill someone in an armed conflict must be taken after a value judgment (which a machine is obviously unable to make and must be made by a human being) is made about that person. Whether a person may be targeted in an armed conflict is dependent on their status (combatant/civilian) and/or the objective impression resulting from their conduct (direct participation in hostilities).").

[115] ICRC INTERPRETIVE GUIDANCE, *supra* note 110, at 49–57.

[116] *Id.* at 46.

[117] *Id.* at 59.

[118] *Id.* at 59–60.

game and thus recognize their intentions as harmless, while a fully autonomous weapon might see only a person running toward it and two armed individuals."[119]

This is a challenging scenario, as the principle of distinction does not prohibit attacking a child who is directly participating in hostilities.[120] The critical question is whether a human soldier would be more likely than an autonomous weapon to recognize that the guns were not real. It is unlawful to knowingly target someone holding a toy gun, because pointing a toy gun at a soldier does not qualify as direct participation.

It is difficult to believe that an AWS would be more likely than a human soldier to mistakenly conclude that the guns were real. Intent, here, is irrelevant: as Marco Sassóli points out, "[e]ven if the mother was inciting the children to hate and the children were crying out in hate and subjectively willing to kill the soldier, the latter could not fire if it was apparent that the pistols were toy guns."[121] Instead, the issue is solely object recognition: are the guns real or toys? This is the kind of task for which machines are particularly well-equipped. Patriot One's "PatScan" threat-detection product, for example, consists of "algorithms that have been trained to recognize weapons from the signatures we get through . . . video object recognition."[122] The system is able to recognize certain weapon types, such as semi-automatic assault rifles, with nearly 95% certainty.[123] By contrast, as discussed in more detail below, humans are prone to a number of cognitive errors that make weapon recognition quite difficult.

Let us assume, however, that the toy guns are indistinguishable from real ones—a particularly convincing AK-47, for example. This is a far more difficult object-recognition task, one that is likely beyond the skills of both humans and machines. The critical question now is whether a human soldier would be more likely than an AWS to infer from the mother's yells and the children's actions that, despite appearing to be holding real guns, the children do not pose an actual threat of harm.

Human Rights Watch does not specify what the frightened mother yells to her children, but it is probably something like "stop playing near the soldier" or "put your toys down." If so, her words are the best indication that the children do not pose a threat, despite the children playing with realistic guns near the soldier. That is a language-recognition task, and it is unlikely that a human soldier would be more capable of recognizing the meaning of the sentence than an autonomous

[119] HUMAN RIGHTS WATCH, *supra* note 88, at 31–32.
[120] *See* ICRC INTERPRETIVE GUIDANCE, *supra* note 110, at 60 (noting that, if they DPH, "[e]ven . . . children below the lawful recruitment age may lose protection against direct attack").
[121] Sassóli, *supra* note 90, at 333.
[122] *PATSCAN Platform Detects Hidden Weapons, Chemicals and Bombs*, TECHREPUBLIC, https://www.techrepublic.com/videos/patscan-platform-detects-hidden-weapons-chemicals-and-bombs [https://perma.cc/P36Q-QEM8] (last visited Oct. 22, 2023).
[123] Winter, *supra* note 94, at 865.

weapon. In fact, if the mother is speaking a foreign language, the AWS would have distinct advantages over the human soldier. It is easier for machines to learn a foreign language than humans; machine translation is much faster;[124] and machine hearing will eventually be far more sensitive than human hearing—if it isn't already.[125]

That said, if the children have realistic guns and their mother does not yell an unambiguous word like "play" or "toy," the ability to accurately determine whether the children pose a threat to the soldier may well require, following Sharkey, the kind of "understanding of other people's intentions and their likely behaviour"[126] that machines lack. An autonomous weapon might be equally able to recognize the mother's look as one of fear and not anger, but a human soldier would be better able to understand the *meaning* of the mother's look of fear—that it is more likely to be motivated by the possibility of a terrible misunderstanding than by a desire to prevent harm to two young insurgents. And there is no question that a human soldier would be better able to recognize the children's actions, despite their seemingly real guns, as play instead of as an attack.

In short, there will indeed be situations in armed conflict—particularly involving direct participation in hostilities—where the uniquely human ability to discern a potential target's intent will determine whether a combatant is able to comply with the principle of distinction. As this section has shown, however, that ability is much less important to the principle than critics of autonomous weapons acknowledge. In most combat situations, the intent of the target will be irrelevant to determining his targetability.

B. Proportionality

The principle of proportionality, codified in Article 51(5)(b) of AP I,[127] prohibits "an attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated." Ensuring that an otherwise legitimate attack is not disproportionate thus requires the attacker to make three different assessments: the military advantage the attack will gain; the incidental civilian damage the attack will cause; and whether the latter will be excessive in relation to the former.

[124] *See, e.g.*, Fouad Habash, *AI Translation vs. Human Translation: Pros and Cons*, BLEND (Dec. 14, 2022), https://www.getblend.com/blog/ai-translation-vs-human-translation-pros-and-cons/ [https://perma.cc/YGP7-DTJU] ("AI translation can deliver near-instant results, making it ideal for when you need to meet an urgent deadline.").

[125] *See generally* RICHARD F. LYON, HUMAN AND MACHINE HEARING: EXTRACTING MEANING FROM SOUND (2017).

[126] Sharkey, *supra* note 86, at 379.

[127] "[T]he following types of attacks are to be considered as indiscriminate . . . [A]n attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."

Critics of autonomous weapons argue that human understanding is even more essential to the principle of proportionality than it is to the principle of distinction,[128] because all three proportionality assessments are inherently fact-sensitive and context-dependent. According to Asaro, for example, the need "for a human being to make an informed decision" is particularly acute "in proportionality decisions in which one must weigh the value of human lives, civilian and combatant, against the values of military objectives. None of these are fixed values, and in some ways these values are set by the very moral determinations that go into making proportionality judgements."[129] Similarly, Jarna Petman doubts that AWS will ever be able to respect the principle of proportionality because the principle "requires a subjective assessment. The practical application thereof requires a weighing of potentially competing interests: military advantage and the protection of civilians. This weighing of interests is only possible on a case-by-case basis: different circumstances require different responses."[130]

In some environments, critics are right that autonomous weapons will struggle to comply with the principle of proportionality. Although sophisticated methodologies already exist for assessing the anticipated incidental damage of a specific attack[131] that could be programmed into an AWS,[132] assessing military advantage is more complicated. Even scholars who support the use of autonomous weapons accept that, to quote Schmitt, "[g]iven the complexity and fluidity of the modern battlespace, it is unlikely in the near future that, despite impressive advances in artificial intelligence, 'machines' will be programmable to perform robust assessments of a strike's likely military advantage."[133] Moreover, machines would find it particularly difficult to assess whether the anticipated incidental

[128] *See, e.g.*, HUMAN RIGHTS WATCH, *supra* note 21, at 8 ("The obstacles presented by the principle of distinction are compounded when it comes to proportionality."); JARNA PETMAN, AUTONOMOUS WEAPONS SYSTEMS AND INTERNATIONAL HUMANITARIAN LAW: "OUT OF THE LOOP"? 37 (2018) (arguing that "the question of whether an autonomous system could undertake a proportionality analysis may be even more fraught than the one concerning the ability of AWS to comply with the principle of distinction").

[129] Peter Asaro, *On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making*, 94 INT'L. REV. RED CROSS 687, 701 (2012).

[130] PETMAN, *supra* note 128, at 36; *see also* AMOROSO ET AL., *supra* note 18, at 24 ("The prospect of developing AWS capable of assessing proportionality with sufficient competence prior to and during an attack appears at the present to be groundless, insofar as proportionality analysis, like distinction, relies heavily on qualitative elements and open-textured standards such as the judgment of a 'reasonable military commander'."); Noel Sharkey, *Cassandra or False Prophet of Doom: AI Robots and War*, 23 IEEE INTELL. SYST. 14, 17 (2008) (claiming that the decision "requires human judgment" because "there is no sensing capability that would help a robot make such a determination").

[131] *See* Elliot Winter, *The Compatibility of Autonomous Weapons with the Principles of International Humanitarian Law*, 27 J. CONFLICT AND SEC. L. 1, 16 (2022) (discussing "collateral damage estimation methodologies" currently used by various militaries).

[132] Schmitt, *supra* note 108, at 20.

[133] *Id.* at 21; *see also* Sassóli, *supra* note 90, at 331 ("In my view, the greatest difficulty an autonomous weapon system will have in applying the proportionality principle is not linked to the evaluation of the risks for civilians and civilian objects, but to the evaluation of the military advantage anticipated.").

damage of an attack would be "excessive" in relation to its expected military advantage, because—contrary to Winter's suggestion[134]—it is not always possible to quantify military advantage and collateral damage in a way that allows them to be mathematically compared. Such quantification is not only inherently value-laden;[135] "often the expected military advantage and collateral damage are dissimilar, for example, when the expected destruction of vital civilian infrastructure needs to be compared to the destruction of one enemy tank with four enemy soldiers."[136]

These problems, however, are not unique to autonomous weapons. On the contrary, there is no reason to believe that human soldiers are any better at making the inherently fact-sensitive and context-dependent assessments that the principle of proportionality requires, especially given the "significant ambiguity surrounding the standard's actual content."[137] This is particularly true of soldiers engaged in remote warfare: much can change in terms of civilian presence between the time a ship fires a cruise missile or a bomber releases its payload and the time the weapon reaches its military target—up to 15 minutes for the cruise missile.[138] But it is also true of soldiers present on the battlefield, who will often be "confronted with unexpected or confusing events when making a time sensitive decision in combat."[139] It is not surprising, therefore, that "many military lawyers have questioned whether human soldiers are capable of truly applying this ambiguous test either."[140] Militaries simply assume that they can accurately apply the proportionality test, despite the evidence to the contrary.[141]

[134] Winter, *supra* note 131, at 17 (arguing that "the solution to the incommensurability problem is to calculate collateral damage in terms of lives that will be lost or injuries that will be caused . . . and then to calculate military gain in terms of the lives that will be saved or the injuries that will be prevented . . . These two values can then be compared on a like-for-like basis to render a robust proportionality assessment").

[135] *See, e.g.*, PETMAN, *supra* note 128, at 36–37 ("The proportionality analysis is too highly contextual to allow for it to be reduced, for example, to a rule that you can have one civilian casualty per one combatant killed; or two civilian casualties per a unit commander; or three civilian casualties per one tank destroyed.").

[136] Jeroen van den Boogaard, *Proportionality and Autonomous Weapons Systems*, 6 J. INT'L HUMANITARIAN LEGAL STUD. 247, 268 (2015).

[137] Nehal Bhuta, Susanne Beck & Robin Geiss, *Present Futures: Concluding Reflections and Open Questions on Autonomous Weapons Systems*, *in* AUTONOMOUS WEAPONS SYSTEMS: LAW, ETHICS, POLICY 347, 372 (Nehal Bhuta, Susanne Beck & Robin Geiss eds., 2016).

[138] Birnbacher, *supra* note 58, at 119 (noting that "judgments of proportionality are no less difficult in air strikes and long-range attacks than they are with AWS").

[139] Schmitt, *supra* note 108, at 21.

[140] Kenneth Anderson, Daniel Reisner & Matthew C. Waxman, *Adapting the Law of Armed Conflict to Autonomous Weapon Systems*, 90 INT'L L. STUD. 386, 402 (2014).

[141] *Cf.* George R. Lucas, *Industrial Challenges of Military Robotics*, 10 J. MIL. ETHICS 274, 183 (2011) ("LOAC is written so as to 'offload' most of the troubling questions onto the practitioners and hold them responsible for our post facto review of their decisionmaking under stringent circumstances in the field. That is pretty shoddy legal guidance, and highly questionable from a moral point of view, but it is how things work in the human case.").

It is also important not to overstate the limits of autonomous weapons. To begin with, as Kenneth Anderson and Matthew C. Waxman point out, AWS can be used regardless of their ability to assess proportionality in operational environments where they will be used "only for 'machine-on-machine' encounters, such as missile defense," or where "there are few if any civilians present," such as "an attack against an undersea submarine."[142] Moreover, even in contested operational environments like urban warfare, situations will arise in which the difficulty of assessing proportionality would not prevent an autonomous weapon from launching an attack. For example, none of the proportionality assessments would be difficult for a machine tasked with destroying a particularly high-value target (a bin Laden, say). Given the significant military advantage that killing the target would create, only massive incidental damage would be considered excessive.[143]

Most important of all, though, autonomous weapons have one significant advantage over human soldiers in terms of proportionality: targeting accuracy. The greater the precision of an attack, the less collateral damage it will produce, making it easier to determine *ex ante* whether the attack will likely be disproportionate. Attacks launched by highly accurate AWS will thus create fewer and easier proportionality issues than less accurate weapons like long-range artillery or high-altitude bombers.[144] Consider, for example, the scenario that opens Future of Life's short film "Slaughterbots," in which the Steve Jobs-like figure kills a target by using an autonomous drone smaller than his hand to fire "three grams of shaped explosive" into the target's forehead.[145] It is not unreasonable to be concerned, as Amoroso et al. are, that tiny yet lethal autonomous weapons will provide militaries with "entirely new means for assassinations."[146] Yet as long as such AWS are directed at military objectives—such as the "military leadership or sensitive military infrastructure" Amoroso et al. mention[147]—their use will always be proportionate, because the expected incidental civilian damage will be zero.

[142] ANDERSON & WAXMAN, *supra* note 6, at 6; *see also* Corn, *supra* note 93, at 231 ("The impact of this concern may very well turn on the nature of the weapon and will certainly be impacted by the established parameters of permissible use. For example, this would be a minimal concern for a weapon authorized for use only in areas with minimal to no civilian presence, whereas authorizing use of the weapon in a civilian population centre would require a very different cognitive capacity.").

[143] *Cf.* Jeffrey S. Thurnher, *No One at the Controls: Legal Implications of Fully Autonomous Targeting*, 67 JFQ 77, 83 (2012) ("Regardless of geography, LARs might be appropriate when the target is one of particularly high value. In such situations, a commander may have fewer proportionality concerns or might at least be able to quantify the amount of acceptable collateral damage.").

[144] *Cf.* KRISHNAN, *supra* note 20, at 93 (noting that "if the robot used a highly precise microprojectile or a weak focused laser beam, the damage done, even in the case of missing the intended target or choosing a wrong target, would be comparatively small").

[145] Stop Autonomous Weapons, *Slaughterbots*, YOUTUBE (Nov. 12, 2017), https://www.youtube.com/watch?v=9CO6M2HsoIA.

[146] AMOROSO ET AL., *supra* note 18, at 39.

[147] *Id.*

It is true, then, that autonomous weapons may operate in environments where they will find it difficult to comply with the principle of proportionality. But not in all environments—and in the complicated ones, there is no reason to believe that human soldiers will do any better. On the contrary, from a proportionality perspective, the greater targeting accuracy of autonomous weapons will eventually give them a significant advantage over human soldiers in situations like close-up urban combat.

C. The Limits of the Human

As the analysis above indicates, critics of autonomous weapons significantly overstate the extent to which human understanding is necessary for IHL compliance. This is particularly true for the principle of distinction because some of its central requirements, such as identifying combatants and recognizing surrender, involve little more than object recognition—the kind of task that machines already do particularly well and will only get better at. The principle of proportionality is more complicated, because the three assessments it requires are fact-specific and context-dependent. But those assessments are often so subjective and so dependent on comparing the incommensurable that they defy even human judgment, and there are aspects of the assessments, such as the calculation of expected incidental damage, that favor machines.

None of this means that human understanding is unimportant. As the example of children playing with realistic toy guns indicates, there will indeed be situations in which soldiers are unlikely to comply with IHL unless they are human. But that acknowledgment comes with a very important caveat: there is a fundamental difference between humans *having the ability to make accurate decisions* and humans *actually making them*.

Consider the scenario just mentioned. For the soldier to conclude that the children are not a threat and hold his fire, he must (1) accurately determine that the mother's look is one of fear, not of anger; (2) correctly determine that the children are playing, not threatening him; (3) properly infer from those two circumstantial facts that the children are not directly participating in hostilities; and (4) lower his own weapon before opening fire in self-defense. In perfect circumstances, a soldier might be able to complete those cognitive tasks. But there are no perfect circumstances in combat, so we need to make the scenario more realistic. Assume that the soldier is a new recruit, hasn't slept for thirty-six hours, is terrified of being killed, and recently saw one of his closest friends from basic training die at the hands of the enemy. Assume that he comes across the children while noisy, hot, and dangerous fighting is raging around him. And assume that the boys are fifteen years old—still children, but more than old enough to pose a mortal threat. This is a very different situation, to put it mildly. In fact, cognitive psychology tells us that the soldier is now very likely to make a serious mistake—misperceiving the mother's emotion as anger instead of fear, misjudging the children pointing their real-looking toy guns at him as a hostile act instead of as happy play, or simply pulling the trigger reflexively instead of deliberately releasing it. If the soldier makes any of those mistakes, the children are dead and the soldier has violated the principle of distinction.

The point of revising the scenario is this: when critics claim that autonomous weapons will never be able to comply with IHL as well as human soldiers, they are implicitly comparing machines to a very specific and highly idealized human—one who normally perceives the world accurately, understands rationally, is able to quarantine negative emotions, and reliably translates thought into action.[148] As the rest of this section demonstrates, however, that ideal human does not exist. On the contrary, decades of psychological research indicates that human decision-making is profoundly distorted by cognitive and social biases, by physiological limits, by situational constraints, and by negative emotions. And that is particularly true when humans find themselves in dangerous and uncertain situations like combat.

1. Cognitive Biases

Theories of how humans make decisions generally distinguish between two types of thinking, Type 1 and Type 2. Type 1 thinking is "generally intuitive and automatic, sub-conscious, associative, affective and heuristic-based."[149] Type 2 thinking, by contrast, "entails deliberate and controlled processes, and is slow, effortful, conscious, and rule-based."[150]

Cognitive psychologists believe that humans make almost 95% of their decisions using heuristic-based Type 1 thinking.[151] The percentage is likely even higher for soldiers during combat, where the speed of a decision can determine

[148] *See, e.g.*, Connal Parsley, *Automating Authority: The Human and Automation in Legal Discourse on the Meaningful Human Control of Lethal Autonomous Weapons Systems*, *in* ROUTLEDGE HANDBOOK OF INTERNATIONAL LAW AND THE HUMANITIES 432, 439 (Shane Chalmers & Sundhya Pahuja eds., 2021) ("It is important to understand . . . the peculiar role of automation within [the self-image of the human as non-automated] . . . Humanity has habitually been attributed a 'higher' element (soul, or reason, for example) that masters and governs its material or animal 'lower' part. This 'something higher' has been invoked at decisive moments to resist increasingly sophisticated accounts of all life as 'mechanical' or 'automated'."); John Williams, *Locating LAWS: Lethal Autonomous Weapons, Epistemic Space, and "Meaningful Human" Control*, 6 J. OF GLOB. SEC. STUD. 1, 9 (2021) (pointing out that "compliance debates about LAWS rest on comparisons to the ideal-type human combatant: the IHL compliant just warrior"); Thomas Gregory, *Dangerous Feelings: Checkpoints and the Perception of Hostile Intent*, 50 SEC. DIALOGUE 131, 137 (2019) ("[T]he importance of feelings, intuitions and affects is strangely absent from debates about hostile intent, which tend to assume that soldiers are fully rational subjects who consciously apply the rules of engagement to the particular dilemma that confronts them.").

[149] KAREL VAN DEN BOSCH & ADELBERT BRONKHORST, HUMAN-AI COOPERATION TO BENEFIT MILITARY DECISION MAKING 2 (2018) ("A heuristic is basically a rule of thumb that provides a solution of a complex problem by simplification.").

[150] *Id.*

[151] *See, e.g.*, Pat Croskerry, Geeta Singhal & Sílvia Mamede, *Cognitive Debiasing 1: Origins of Bias and Theory of Debiasing*, 22 BMJ QUAL. SAF. ii58, ii58 (2013).

survival and situations are marked by great complexity and uncertainty.[152] The problem with Type 1 thinking is that it is much less accurate than Type 2: although heuristics are useful for managing time and information constraints, they are particularly prone to errors of judgment known as "cognitive biases."[153] Cognitive biases lead humans using Type 1 thinking to make decisions that "systematically deviate from logic or utility."[154]

A comprehensive discussion of the cognitive biases that may lead human soldiers to make irrational decisions in combat is beyond the scope of this article, because researchers have identified more than ninety-five biases that distort human judgment.[155] It is nevertheless worth exploring some of the most important ones.

a. Stereotyping

As noted above, although the ability to determine a target's intent is less important to IHL compliance than critics of autonomous weapons acknowledge, there are some situations—such as the playing children scenario—in which that ability may well make the difference between compliance and tragedy. Critics are also correct to insist that machines will never be able to mindread, as the ability is often called in cognitive-psychological literature,[156] in the way that humans do.

But that does not mean *humans* generally engage in accurate mindreading. Critics of autonomous weapons simply assume that they do. Some scholars are explicit about that assumption. Sparrow, for example, claims that:

> Human beings have a tremendously sophisticated and powerful capacity to interpret the actions of other human beings and to identify their intentions—to 'read minds'—which has been honed by millennia of primate evolution wherein the ability to know what

[152] *See* JOSEPH RODMAN, COGNITIVE BIASES AND DECISION MAKING: A LITERATURE REVIEW AND DISCUSSION OF IMPLICATIONS FOR THE U.S. ARMY 7 (2015) ("Research suggests that the consequences of intuitive decision making, and therefore of relying on heuristics and succumbing to cognitive biases, becomes more prevalent in situations of greater complexity or uncertainty."); *see also* James Kwoun, *Cognitive Biases and the Need for Analytic Tradecraft Standards in Large-Scale Ground Combat Operations*, 47 MIL. INTEL. 40, 40 (2021) ("Studies have shown that these biases become more likely under ambiguous, traumatic, and time-constrained circumstances, which are exactly the challenges analysts will encounter during a large-scale ground combat operations environment.").

[153] Croskerry, Singhal & Mamede, *supra* note 151, at ii58.

[154] *Id.*; *see also* Blair S. Williams, *Heuristics and Biases in Military Decision Making*, MIL. REV. 40, 41 (2010) (noting that "[i]n the course of these mental processes of simplifying an otherwise overwhelming amount of information, we regularly inject cognitive bias. Cognitive bias comes from the unconscious errors generated by our mental simplification methods").

[155] *List of Cognitive Biases and Heuristics*, THE DECISION LAB, https://thedecisionlab.com/biases [https://perma.cc/L5TR-KRM9] (last visited Oct. 14, 2023).

[156] *See, e.g.*, Guarini & Bello, *supra* note 89, at 131 ("There is a significant literature in cognitive science and philosophy on mental state attribution (sometimes referred to as 'theory of mind,' or 'mentalizing,' or 'mindreading,' with nothing psychic intended).").

> other individuals were thinking and were about to do provided a crucial selective advantage.[157]

More often, though, the assumption remains implicit in the critique of AWS. For example, when Human Rights Watch claims that "fully autonomous weapons would not possess human qualities necessary to assess an individual's intentions" because "[o]ne way to determine intention is to understand an individual's emotional state, something that can only be done if the soldier has emotions,"[158] the organization implies that human soldiers are able to accurately assess a target's intentions as they, unlike killer robots, do have emotions.

Like so many critiques of autonomous weapons, this valorization of the human imagines face-to-face combat. Even if humans can mindread in a way machines cannot, that ability is irrelevant in remote warfare. The bombardier and the artillery operator never see the people they kill.[159] And although the UAV operator might "see" his target, the visual limitations of drones mean that he will not have access to the kind of behavioral cues that successful mindreading requires.[160]

Even in face-to-face combat, however, human soldiers are unlikely to accurately mindread their targets. AWS critics emphasize ambiguous situations on the battlefield, where the target's behavior does not make clear whether he is a combatant, civilian, or a civilian directly participating in hostilities.[161] When an individual is faced with an ambiguous situation, research indicates that his mindreading strategy—his method for determining the target's intent—will be determined by how similar he perceives the target to be to himself:

[157] Sparrow, supra note 83, at 707; *see also* Sharkey, *supra* note 8649, at 379 ("In a war with non-uniformed combatants, knowing who to kill would have to be based on situational awareness and on having human understanding of other people's intentions and their likely behaviour. In other words, human inference is required. Humans understand one another in a way that machines cannot.").

[158] HUMAN RIGHTS WATCH, *supra* note 88, at 31; *see also* Guarini & Bello, *supra* note 89, at 138 ("A system without emotion (or at least some sort of proto-emotional functional counterpart of emotion) could not predict the emotions or action of others based on its own states because it has no such emotional states.").

[159] *Cf.* Schmitt & Thurnher, *supra* note 6, at 248 ("[W]hile it may be true that human perception of human activity can sometimes enhance identification, human-operated systems already engage targets without the benefit of emotional sensitivity. For example, human-operated 'beyond visual range' attacks are commonplace in modern warfare.").

[160] See, for example, the photographs of drone footage in Joanna Tidy, *Visual Regimes and the Politics of War Experience: Rewriting War "from Above" in WikiLeaks' "Collateral Murder"*, 43 REV. INT. STUD. 95, 97 (2016).

[161] *See, e.g.*, Wagner, *supra* note 112, at 1392 ("Not only would AWS have to be able to distinguish civilians from military personnel, but it must also decide if a civilian is taking a direct part in hostilities. These situations are challenging for humans to judge, and it does not appear that the necessary contextual analysis is amenable to easily programmable quantitative assessments at this time.").

> When perceivers assume higher levels of general similarity to a target group, they engage in higher levels of projection on specific attributes, introspecting about their own attitudes and qualities and ascribing them to the target . . . When perceivers assume lower levels of general similarity to a target, they engage in higher levels of stereotyping, turning to implicit beliefs about what a particular group is like.[162]

The problem with using stereotypes to interpret the intentions behind ambiguous behavior is that stereotypes are often inaccurate. A striking example comes from what are known as "shoot/no-shoot" experiments involving racial stereotypes. In a series of laboratory experiments, Joshua Correll et al. presented police officers with short glimpses of Black and white men holding either guns or common household objects such as mobile phones, cans, and wallets. The officers were asked to decide as quickly as possible whether the suspect was holding a gun and, if so, whether they posed an imminent threat of danger that justified shooting them.[163] The results of the experiments were consistent: officers mistakenly shot Black suspects far more often than they mistakenly shot white ones. The race of the target, however, did not influence the officers' ability to distinguish armed from unarmed suspects (object recognition). Instead, "if a target was African American, participants generally required less certainty that he was, in fact, holding a gun before they decided to shoot him" (decision-making).[164] According to the researchers, the differential treatment reflected the influence of inaccurate stereotypes: "ethnicity influences the shoot/don't shoot decision primarily because traits associated with African-Americans, namely 'violent' or 'dangerous', can act as a schema to influence perceptions of an ambiguously threatening target."[165]

These findings have been replicated in studies involving American military cadets. Kevin K. Fleming et al. primed test subjects by showing them a series of white, Black, and Middle Eastern male faces. They then presented the cadets with brief glimpses of various men in civilian dress holding either guns or common objects such as a drill and asked them to decide as quickly as possible whether to shoot.[166] As the researchers predicted, the cadets "made more false positive errors when tools were primed by images of Middle Eastern males wearing traditional

[162] Daniel R. Ames, *Strategies for Social Inference: A Similarity Contingency Model of Projection and Stereotyping in Attribute Prevalence Estimates*, 87 J. PERS. & SOC. PSYCHOL. 573, 574 (2004); *see also* Russell W. Clement & Joachim Krueger, *Social Categorization Moderates Social Projection*, 38 J. EXPERIMENT. SOC. PSYCHOL 219, 228 (2002) ("[S]elf-referent knowledge serves as a readily accessible anchor for in-group estimates but . . . is suspended for out-group estimates.").

[163] Joshua Correll et al., *The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals.*, 83 J. PERS. & SOC. PSYCHOL 1314, 1325 (2002).

[164] *Id.*

[165] *Id.*

[166] Kevin K. Fleming, Carole L. Bandy & Matthew O. Kimble, *Decisions to Shoot in a Weapon Identification Task: The Influence of Cultural Stereotypes and Perceived Threat on False Positive Errors*, 5 SOC. NEUROSCIENCE 201, 204 (2010).

clothing," indicating "a negative stereotype has emerged toward Middle-Eastern males to the extent that they are depicted wearing traditional robes and turbans congruent with the cultural stereotype."[167]

The relevance of this research to the autonomous weapons debate is evident. Critics assume that humans will be better than machines at reading the ambiguous behavior of potential targets simply by virtue of their humanity. In fact, human mindreading is likely to be distorted by any number of inaccurate stereotypes—such as the idea that Black and Middle Eastern men are more violent and dangerous than their white counterparts—that would be quite likely to come into play when soldiers have to decide in the heat of battle whether a potential target is a combatant, civilian, or a civilian directly participating in hostilities.

Stereotyping also extends well beyond these kinds of characterological assumptions. Recall the children-playing scenario discussed above. One of the reasons Human Rights Watch believes a machine would be more likely to mistakenly attack the children is that "[a] human soldier could identify with the mother's fear and the children's game and thus recognize their intentions."[168] This argument presumes that a human soldier would accurately recognize the mother's facial expression as fear. Perhaps surprisingly, that is a problematic assumption: a significant amount of research indicates not only that humans rely on stereotypes to connect particular facial expressions to particular emotions,[169] but also that those stereotypes differ significantly between individuals.[170] Those differences are critical, because if individuals have different stereotypes of what a particular emotion looks like, by definition some of them will mindread incorrectly when they apply their stereotypes to a target's facial expression. And indeed, research indicates that test subjects often assume that fear is sadness and sadness is anger.[171] Confusing sadness for anger could be tragic on the battlefield if a soldier relies primarily on facial expression to determine whether an individual poses a threat.

b. Availability Bias

Availability bias occurs "when people judge the likelihood of something happening by how easily they can retrieve similar examples to mind. If an outcome

[167] *Id.* at 217.

[168] HUMAN RIGHTS WATCH, supra note 88, at 31.

[169] *See, e.g.*, Nicola Binetti et al., *Genetic Algorithms Reveal Profound Individual Differences in Emotion Recognition*, 119 PROC. NATL. ACAD. SCI. U.S.A. 1, 2 (2022) (noting, with regard to a variety of expression-recognition tasks, that performance "largely varies as a function of the similarity between an individual's preferred depiction of a facial expression and the test images of facial expressions that are used in a task").

[170] *See, e.g.*, *id.* at 5–6 ("[T]his approach reveals large individual differences in preferred expressions of core emotion categories, with significant overlap between fear and sad categories. These differences in preferred expressions in turn influence emotion recognition, with individual differences in performance explained by the extent to which test stimuli resemble participants' preferred expressions.").

[171] *See* Binetti et al., *supra* note 169, at 6.

is vividly imaginable, the probability of its occurrence is likely to be overestimated."[172] The availability bias is quite likely to distort decision-making when soldiers make split-second targeting decisions, because previous negative experiences in combat may predispose soldiers to overestimate the likelihood that a target's ambiguous behavior is actually threatening.[173] A soldier who has survived an ambush launched by an individual who removed an AK-47 from beneath civilian robes, for example, will be more likely than a soldier who has no such experience to assume that future individuals dressed in civilian robes are planning to ambush him. Similarly, "the subjective probability assessment of future improvised explosive device (IED) attacks will most likely be higher from a lieutenant who witnessed such attacks than one who read about them in situation reports."[174]

c. Imaginability Bias

The imaginability bias is a corollary to the availability bias, applying when an individual has no readily available examples of similar conduct to draw on when trying to interpret ambiguous behavior. In such situations, probability becomes a function of imaginability: the easier it is to imagine a particular course of events, the more probable that course of events will appear to be.[175] That is a problematic heuristic, because there is no necessary correlation between imaginability and the actual likelihood that a particular event will occur.[176] Consider close-up urban combat, the type of combat that involves the most challenging discrimination issues—ones that, according to critics of autonomous weapons, require human judgment. An inexperienced soldier experiencing close-up urban combat for the first time is very likely to imagine scenarios in which he is ambushed by civilians engaging in seemingly innocuous acts such as riding a motorbike or watching him from a rooftop. The mere act of imagining those scenarios will lead the soldier to overestimate the likelihood that a civilian he encounters engaged in one of those activities will ambush him.

d. Base Rate Bias

According to base rate bias, "[w]hen provided with both individuating information, which is specific to a certain person or event, and base rate

[172] R. J. Knighton, *The Psychology of Risk and its Role in Military Decision-Making*, 4 DEF. STUD. 309, 321 (2004).

[173] Research indicates that the availability bias is promoted by time pressure and by ambiguous behavior. BARBARA D. ADAMS ET AL., HUMAN DECISION-MAKING BIASES 12 (2009).

[174] Williams, *supra* note 154, at 60.

[175] *See, e.g.*, R. Scott Rodgers, *Improving Analysis: Dealing with Information Processing Errors*, 19 INT'L J. OF INTEL. & COUNTERINTEL. 622, 632 (2006) ("Biases of imaginability refer to the tendency to retrieve information that is plausible without regard to its probability. People regularly construct a series of possible behaviors or plans based, to a large extent, on their ability to imagine their occurring. By imagining a particular course of events, the likelihood is that they will plan accordingly, regardless of the probability of these events transpiring.").

[176] *See id.* (noting that "[b]eing able to imagine that a client could commit suicide greatly increases a clinician's assessment that it would occur even though it may be extremely unlikely").

information, which is objective, statistical information, we tend to assign greater value to the specific information and often ignore the base rate information altogether."[177] Consider a hypothetical scenario provided by the RAND Corporation to illustrate how base rate bias can lead to tragic targeting decisions. In the scenario, a Joint Force Air Component Commander (JFACC) trying to kill three enemy government officials "receives credible intelligence that three men in tan uniforms are in a white Jeep with a black roof, on the highway heading to the border."[178] Because a decision-support system indicates that the government officials are likely using a white Jeep with a black roof, the commander deploys a UAV to destroy the Jeep, which turns out to be carrying civilians. That mistake is due in large part to base rate bias: the JFACC "has ignored (or not sought out) the base-rate data—that *most* of the cars in the area match the description in the intelligence."[179]

e. Anchoring Bias

Anchoring bias, sometimes referred to as information-order bias, "causes us to rely heavily on the first piece of information we are given about a topic. When we are setting plans or making estimates about something, we interpret newer information from the reference point of our anchor instead of seeing it objectively."[180] Anchoring bias can easily distort targeting decisions, as indicated by a series of experiments involving operators of the Army's Patriot Air Defense System. The operators were asked to decide whether to attack aircraft that appeared one at a time on their screens and engaged in a series of moves that were consistent with either friend or foe status. The researchers predicted that when there was "conflicting information about the aircraft . . . the same information will result in different judgments simply depending on the sequence (or order) in which the information is presented to the operators."[181] That is precisely what they found: the number of errors committed by the operators—attacking friends or not attacking foes—depended on when in the aircraft's track the misleading information was presented: the earlier the information appeared in the track, the more likely it was the operator would rely on it.[182] In other words, because of anchoring bias, the operators suffered from a primacy effect[183] in which earlier inaccurate information

[177] *Base Rate Fallacy*, THE DECISION LAB, https://thedecisionlab.com/biases/base-rate-fallacy [https://perma.cc/AG46-Y43X] (last visited Oct. 14, 2023).
[178] PAUL K. DAVIS, JONATHAN KULICK & MICHAEL EGNER, IMPLICATIONS OF MODERN DECISION SCIENCE FOR MILITARY DECISION-SUPPORT SYSTEMS 102 (2005).
[179] *Id.*
[180] *Anchoring Bias*, THE DECISION LAB, https://thedecisionlab.com/biases/anchoring-bias [https://perma.cc/9YKZ-CLCN] (last visited Oct. 22, 2023).
[181] LEONARD ADELMAN & TERRY A. BRESNICK, EXAMINING THE EFFECT OF INFORMATION SEQUENCE 1 (1995).
[182] *Id.* at 16.
[183] Christopher D. Wickens et al., *The Anchoring Heuristic in Intelligence Integration: A Bias in Need of De-Biasing*, 54 PROC. HUM. FACTORS & ERGONOMICS SOC. ANN. MEETING 2324, 2324 (2010) ("In the case of anchoring, when cues relative to an intelligence assessment arrive over

served to "anchor" their identifications in such a way that later-presented accurate information did not affect their final decision whether to attack.

f. Object Use Bias

Another type of cognitive error that could lead human soldiers to make distinction errors is what might be called "object use bias." Hypothesizing that an individual's perception of objects held by others could be affected by the nature of the objects the individual carried himself, Jessica Witt and James Brockmole conducted a series of experiments in which students were asked to identify as quickly as possible whether men in various images were holding a gun or a neutral object.[184] In some tests, the students held a gun while making the identifications; in others, they held a ball. The results were unequivocal:

> The familiar saying goes that when you hold a hammer, everything looks like a nail. The apparent harmlessness of this expression fades when one considers what happens when a person holds a gun. We have shown here that, having the opportunity to use a gun, a perceiver is more likely to classify objects held by others as guns and, as a result, to engage in threat-induced behavior (in this case, raising a firearm to shoot).[185]

This research is directly relevant to combat. Because soldiers carry weapons, object-use bias indicates that they will be more likely to misperceive objects held by individuals they encounter on the battlefield than they would be if they were unarmed. Moreover, Witt and Brockmole note that their results support the idea that the *mere planning* of "an action with a given object or tool should bias observers to identify similar objects."[186] So even the very nature of combat itself—in which the use of weapons is always expected—is likely to give rise to identification errors.

g. Confirmation Bias

Confirmation bias, often colloquially referred to as "tunnel vision," refers to the tendency of humans "to actively seek information and assign greater value to evidence confirming their existing beliefs rather than entertaining new ones."[187] Confirmation bias, which is particularly likely to occur when "concrete task-

time (or are processed sequentially by the human operator), there is a tendency for the human to give greater weight to the first arriving cue (or cues). That is, 'first impressions are lasting'.").

[184] Jessica K. Witt & James R. Brockmole, *Action Alters Object Identification: Wielding a Gun Increases the Bias to See Guns*, 38 J. OF EXPER. PSYCHOL.: HUM. PERCEPTION & PERFORMANCE 1159, 1160 (2012).

[185] *Id.* at 1165.

[186] *Id.* at 1166.

[187] *Confirmation Bias*, THE DECISION LAB, https://thedecisionlab.com/biases/confirmation-bias [https://perma.cc/6KUU-JTSA] (last visited Oct. 15, 2023).

specific information is lacking" and the individual is stressed and under time pressure,[188] significantly increases the likelihood of mistaken judgment when a person's existing beliefs are incorrect.

These factors suggest that soldiers are very likely to suffer from confirmation bias when they attempt to interpret ambiguous behavior on the battlefield. Indeed, a number of high-profile tragedies have been explicitly attributed to soldiers ignoring or discounting evidence that contradicted their initial belief that particular objects were targetable. According to Chien Wen Chia, for example, the best explanation of why the *USS Vincennes* mistakenly shot down an Iranian passenger jet, killing all 290 people on board, is that the captain's "tentatively held hypothesis of the approaching aircraft being hostile" led him to ignore the possibility that it was civilian "even though that alternate hypothesis was floated on more than one occasions [sic] in the minds of the crew from its initial detection until its final engagement."[189] An Air Force general who investigated a 2021 UAV attack on a car in Afghanistan that killed 10 civilians, including seven children, reached a similar conclusion.[190] Specifically citing confirmation bias, the general attributed the failure to recognize that civilians were in the car to the military expecting an ISIS suicide attack to come from a car of that make and model in the same location. Because the military assumed the civilian car was the car that ISIS intended to detonate, UAV operators ignored evidence inconsistent with that assumption—such as the presence of a California-based nutrition NGO in that location and the operators seeing people moving around unexpectedly in the compound.[191]

Confirmation bias, it is important to note, often combines with other cognitive biases to distort decision-making even further. For example, when an individual assumes that he can interpret a new situation because he has experienced an ostensibly similar situation in the past—availability bias—he is quite likely to ignore aspects of the new situation that are inconsistent with his assumption.[192] Similarly, relying on imaginability to assess the likelihood of a particular event occurring—imaginability bias—is not only likely to be inaccurate, it "can also result in premature cognitive closure, which makes the decision-maker insensitive

[188] ADAMS ET AL., *supra* note 173, at 22.

[189] CHIEN WEN CHIA, COUNTERING POSITIVE CONFIRMATION BIASES IN COMMAND TEAMS: AN EXPERIMENT WITH DIFFERENT INTERVENTIONS 4 (2005).

[190] Rose L. Thayer, *10 Civilians, Including 7 Children, Killed in US Drone Strike During Final Days of Afghanistan Pullout, Top General Says*, STARS AND STRIPES, Sept. 17, 2021, https://www.stripes.com/theaters/middle_east/2021-09-17/drone-strike-kabul-afghanistan-civilians-children-killed-isis-2923607.html [https://perma.cc/BLL8-BD8W].

[191] *See id.*

[192] *See* ADAMS ET AL., *supra* note 173, at 12 (noting that "pre-existing templates are not necessarily accurate in each new situation, and this can lead to systematic errors . . . when a suboptimal category is used to make a decision. For example, a display operator may be inclined to disregard a target that shared some characteristics of commonly benign targets. This could lead to a decision error").

to alternative courses of action or information that runs contrary to his view of the world."[193]

Peter Margulies provides a particularly striking example of how confirmation bias can combine with a different cognitive bias to produce a tragic error in judgment: the infamous U.S. military attack on the Medecins Sans Frontieres (MSF) hospital in Afghanistan in 2015, which killed dozens of civilians. After initially assuming that the hospital was a Taliban base, the military ignored extensive indications to the contrary because it relied on the hospital building's use of a distinctive type of arch—one that matched a nearby Taliban base—to confirm its initial belief. "In reality, since arches are a common feature of buildings in the region, the presence of arches on the MSF facility was neutral information that did not prove or disprove their initial targeting theory."[194] Margulies describes the mistaken attack as demonstrating confirmation bias, which it does. But it is also an example of base-rate bias, because the military assumed that a particular arch implied Taliban presence without knowing how many non-Taliban buildings in the area used the same one.

2. Physiological Limits

The ability of soldiers to perform adequately and make accurate decisions in combat is also affected by human physiological limits. Some of those limits are "merely" physical, such as those that affect marksmanship—a critical ability for IHL compliance. A member of the infantry in the U.S., for example, must be able to hit 36 of 40 (90%) targets to qualify as an "expert," 30 of 40 (75%) to qualify as a "sharpshooter," and 23 of 40 (58%) to qualify as a "marksman."[195] An Army study of 2,000 soldiers found that, in a simulation approximating combat, only 7% were experts, 31% were sharpshooters, and 37% were marksmen.[196] That means 25% of soldiers scheduled for deployment were not able to hit even 6 out of 10 targets—and 62% couldn't hit 3 out of 4. Those are disturbing statistics—and yet they compare favorably to police officers, given that a RAND study found that, between 1998 and 2006, NYPD officers who were shot at and returned fire hit the target anywhere on the body only 18% of the time.[197] In other words, 4 out of 5 bullets fired by the police officers missed their targets completely.

Although the Army marksmanship simulation included real-world combat aspects such as a malfunctioning weapon and the need to reload, it did not ask soldiers to shoot while physically exhausted—an incredibly common situation in

[193] Knighton, *supra* note 172, at 322.

[194] Peter Margulies, *The Other Side of Autonomous Weapons: Using Artificial Intelligence to Enhance IHL Compliance*, THE IMPACT OF EMERGING TECHNOLOGIES ON THE LAW OF ARMED CONFLICT 147, 150 (Ronald T.P. Alcala & Talbot Jensen eds., 2019).

[195] Chris Jenks & Heather Roff, *Is (or Should) Better Be Good Enough? Legal Reviews of Learning or Adaptive Weapons Systems* 13 (2023) (unpublished manuscript) (on file with author).

[196] *Id* at 13–14.

[197] *Id.* at 14.

combat.[198] Had it done so, the results would almost certainly have been even worse, because physical exhaustion significantly degrades concentration and reaction time.[199] The effects of exhaustion, moreover, are not limited to such mechanical activities. On the contrary, exhaustion also degrades general cognition, situational awareness, moral decision-making, and the ability to respond appropriately to negative emotional stimuli.[200] Cognitive performance on a task involving decision-making and short-term memory, for example, has been shown to decline 25% for every day that an individual goes without sleep.[201] Such deficits can easily lead soldiers to violate IHL, as evidenced by a study in which 59% of Norwegian military cadets who were extremely sleep-deprived were willing to follow an order to shoot at people who could not be positively identified as combatants.[202]

Another important physiological phenomenon that undermines performance and rational decision-making in combat is excessive cognitive load, which refers to "the load that performing a particular task imposes on the cognitive system."[203] Because individuals have a limited working memory capacity, they can only handle so much cognitive load before their cognitive performance declines.[204]

Cognitive overload is a persistent problem in combat of all kinds,[205] but it is particularly acute in urban warfare, because "population-centric warfare increases the unpredictability, uncertainty, complexity, and moral ambiguity of military operations." [206] Simply put, close-up combat imposes intense cognitive demands on soldiers:

> Soldiers often have to perform vigilance type task that are mentally fatiguing such as stationary surveillance (e.g., sentry or security operations) or extended mobile patrols (e.g., monitoring

[198] *See, e.g.*, Elizabeth A. Stanley & Kelsey L. Larsen, *Difficulties with Emotion Regulation in the Contemporary U.S. Armed Forces: Structural Contributors and Potential Solutions*, 47 ARMED FORCES & SOC. 77, 87 (2021) (citing empirical studies which find deployed soldiers sleep 5.5-6.5 hours per night on average).

[199] *See* Eugene Aidman, Simon A. Jackson & Sabina Kleitman, *Effects of Sleep Deprivation on Executive Functioning, Cognitive Abilities, Metacognitive Confidence, and Decision Making*, 33 APPL. COGNITIVE PSYCHOL. 188, 198 (2019).

[200] *See* Stanley & Larsen, *supra* note 198, at 87–89.

[201] Nancy J. Wesensten, Gregory Belenky & Thomas J. Balkin, *Sleep Loss: Implications for Operational Effectiveness and Current Solutions*, *in* 1 MILITARY LIFE: THE PSYCHOLOGY OF SERVING IN PEACE AND COMBAT 81, 84 (Thomas W. Britt, Amy B. Adler & Carl Andrew Castro eds., 2006).

[202] Rolf P. Larsen, *Decision Making by Military Students Under Severe Stress*, 13 MIL. PSYCH. 89, 93 (2001).

[203] John Sweller, Jeroen J.G. van Merrienboer & Fred G. W. C. Paas, *Cognitive Architecture and Instructional Design*, 10 EDUC. PSYCHOL. REV. 251, 266 (1998).

[204] *Id.*

[205] *See, e.g.*, Kristy Martin et al., *Physiological Factors Which Influence Cognitive Performance in Military Personnel*, 62 HUM. FACTORS 93, 93 (2020) ("Military personnel face particularly unique challenges to cognition, including combinations of extreme physical and mental fatigue, high levels of anxiety and stress, and environments of great unpredictability.").

[206] Stanley & Larsen, *supra* note 198, at 88.

> environment in a mounted convoy or on dismounted patrol). Without warning, soldiers may need to engage multiple targets within an urban environment that contains a civilian population. Thus, soldiers need to not only be accurate and fast, but also judicious in their execution of sound judgment prior to pressing trigger.[207]

These cognitive demands are heightened by the fact that, on the modern battlefield, soldiers "are required to process ever-increasing amounts of information as sensors, data handling, and information displays become more prevalent."[208]

Cognitive overload routinely causes soldiers to make distinction mistakes in combat. Studies have consistently shown, for example, that soldiers "show higher error and fratricide rates under dual-task workload conditions."[209] One study found that merely performing a cognitively demanding task for forty-nine minutes—a pale imitation of what actual combat requires—made soldiers 16% more likely to erroneously shoot at a target.[210] That work was later extended by different researchers, who found that the error rate more than doubled (33%) when the soldiers, 96% of whom were rated "expert" or "sharpshooter," spent 49 minutes completing a cognitively demanding task that (unlike in the earlier study) required the same kind of "active response inhibition" as marksmanship.[211] In a third study, "[s]pecifically, the authors found that participants shooting in a target rich environment . . . had significant increases in errors of commission (i.e., incorrectly shooting instead of withholding shot) and speeded responses resulting in speed accuracy trade-offs."[212]

3. Situational Constraints

Cognitive overload becomes even worse—and thus leads to even more mistakes—when soldiers are exposed to normal combat conditions such as noise, heat, and time pressure. Noise significantly degrades performance of combat-related tasks such as disposing explosive ordinance, with soldiers working more quickly and making more errors, and high-noise situations inhibit an individual's

[207] James Head et al., *Prior Mental Fatigue Impairs Marksmanship Decision Performance*, 8 FRONTIERS PHYSIOL. 1, 8 (2017).

[208] David R. Scribner, *Predictors of Shoot–Don't Shoot Decision-Making Performance: An Examination of Cognitive and Emotional Factors*, 10 J. COGNITIVE ENGINEERING & DECISION MAKING 3, 3 (2016); *see also id.* at 2 ("The likelihood of soldiers developing mental fatigue in the field has only increased with the advent of head-up displays and changing battlefield scenarios [e.g., constant scanning for improvised explosive devices (IEDs)].").

[209] Scribner, *supra* note 208, at 3.

[210] Martin et al., *supra* note 205, at 11.

[211] Head et al., *supra* note 207, at 6.

[212] *See id.* at 2

attention to the "interpersonal demands of others," such as needing help.[213] Heat negatively affects soldiers' ability to estimate time and read maps,[214] two skills that play a critical role in both distinction and proportionality judgments. And time pressure increases performance speed while decreasing performance accuracy because of "the cognitive demands, or information overload, imposed by the requirement to process a given amount of information in a limited amount of time."[215]

4. Negative Emotions

Decision-making is profoundly influenced by emotions,[216] particularly in situations like combat[217] that involve ambiguity and uncertain outcomes.[218] That is not inherently problematic, because in some situations emotions can be useful guides to conduct.[219] A vast amount of research indicates, however, that "integral emotion inputs to decision making, especially perceptually vivid ones, can override otherwise rational courses of action."[220]

The impact of combat-related stress on performance and decision-making illustrates this problem. Such stress not only leads to "perceptual tunneling, reduced working memory, and performance rigidity," it also makes soldiers prone to what is known as ballistic decision-making: "making decisions without checking the consequences."[221] In a pair of studies, for example, H.R. Lieberman et al. subjected experienced Army Rangers and Navy SEALs to highly stressful simulated combat situations, such as lack of sleep, extreme weather conditions, demanding physical activities, and verbal abuse by superiors. The results were unambiguous:

[213] James E. Driskell, Eduardo Salas & Joan H. Johnston, *Decision Making and Performance Under Stress*, *in* 1 MILITARY LIFE: THE PSYCHOLOGY OF SERVING IN PEACE AND COMBAT 128, 134 (Thomas W. Britt, Amy B. Adler & Carl Andrew Castro eds., 2006).

[214] *See* Eric R. Muth et al., *Augmented Cognition: Aiding the Soldier in High and Low Workload Environments Through Closed-Loop Human-Machine Interactions, in* MILITARY LIFE, *supra* note 213, at 108, 110.

[215] Driskell, Salas, & Johnston, *supra* note 213, at 135.

[216] *See* Kathleen L. Mosier & Ute Fischer, *The Role of Affect in Naturalistic Decision Making*, 4 J. OF COGNITIVE ENGINEERING & DECISION MAKING 240, 240 (2010); *see also* ADAMS ET AL., *supra* note 173, at 25 ("Humans do not reason without emotion; in fact, affective responses play an important role in memory, decision-making, and judgement, typically in the direction of the salient emotion.").

[217] *See, e.g.*, Siniša Malešević, *Emotions and Warfare: The Social Dynamics of Close-Range Fighting*, OXFORD RESEARCH ENCYCLOPEDIA OF POLITICS 10 (2021) ("Although wars are often conceptualized in instrumentalist and rationalist terms, the actual lived experience of the combat zone is principally defined by variety of emotional reactions. There is no close-range fighting without emotions.").

[218] Mosier & Fischer, *supra* note 216, at 240 ("[I]f a situation is ambiguous with respect to some critical aspect (e.g., certainty of outcome), then incidental emotion (e.g., fear) may impact decision making more (e.g., risk avoidance), as compared with its effect in unambiguous situations.").

[219] *See* Jennifer S. Lerner et al., *Emotion and Decision Making*, 66 ANN. REV. PSYCHOL. 799, 802 (2015).

[220] *Id.* at 803.

[221] Driskell, Salas & Johnston, *supra* note 213, at 140.

> Virtually every task conducted on the battlefield, from the simplest to the most complex, requires the individual to employ multiple cognitive functions. For example, firing a weapon at the right time at the correct target requires the following cognitive elements: vigilance and pattern recognition to detect the target; choice reaction time to fire at the correct target at the right instant; logical reasoning to determine whether firing a weapon at a selected target is tactically appropriate and permitted within the rules of engagement; and short-term memory to insure that the warfighter is aware of the location of friendly forces. The tests we employed measured all of these functions and all were substantially impaired.[222]

"Substantial impairment" is actually an understatement. According to the researchers, "[t]he magnitude of the deficits observed was greater than those produced by alcohol intoxication or treatment with sedating drugs."[223]

Other researchers have reached similar conclusions. In terms of performance, Nicky Nibbeling et al. found that a state of high anxiety reduced the shooting accuracy of experienced infantry by 20 to 40%.[224] In terms of decision-making, Katherine Gamble et al. found using a realistic shoot/no-shoot course that active-duty infantry made more distinction errors ("shot at more targets, friend or foe") in a high-stress situation than in a low-stress one, a bias the researchers attributed to the desire for self-preservation.[225] That result accords with a troubling study conducted by Arne Nieuwenhuys et al. in which police officers were placed in a state of high anxiety and then asked to decide whether to shoot a series of suspects who either had a gun or were trying to surrender. As predicted, in comparison to low-anxiety officers, "the percentage of incorrect responses almost doubled, increasing up to almost 20% . . . implying that in every five cases a surrendering suspect was shot."[226]

Anger, an extremely common battlefield emotion, has similar cognitive effects. In fact, anger is more likely than any other emotion to lead to the kind of Type 1 heuristic-based thinking that often leads to faulty decision-making.[227] Researchers have found, for example, that angry soldiers are "less attuned to the

[222] H.R. Lieberman et al*., The "Fog of War": Documenting Cognitive Decrements Associated with the Stress of Combat, in* PROC. 23RD ARMY SCI. CONF. 1, 6 (2002).

[223] *Id.*

[224] Nicky Nibbeling et al., *The Effects of Anxiety and Exercise-Induced Fatigue on Shooting Accuracy and Cognitive Performance in Infantry Soldiers*, 57 ERGONOMICS 1366, 1377 (2014).

[225] Katherine R. Gamble et al., *Different Profiles of Decision Making and Physiology Under Varying Levels of Stress in Trained Military Personnel*, 131 INT'L J. OF PSYCHOPHYSIOLOGY 73, 78 (2018).

[226] Arne Nieuwenhuys, Geert J. P. Savelsbergh & Raôul R. D. Oudejans, *Shoot or Don't Shoot? Why Police Officers Are More Inclined to Shoot When They Are Anxious*, 12 EMOTION 827, 831 (2012).

[227] Mosier & Fischer, *supra* note 216, at 244 ("Anger, in particular, has been linked with heuristic processing.").

external environment, more likely to ignore new information, more committed to existing preferences, more risk-seeking, and often more reactive."[228] Anger also promotes the use of inaccurate stereotypes to mindread[229] and leads to "carryover of incidental emotions," in which "anger triggered in one situation automatically elicits a motive to blame individuals in other situations even though the targets of such anger have nothing to do with the source of the anger."[230] This phenomenon could obviously lead to distinction errors in combat, as well as to unethical behavior more generally. To offer a specific example of the latter, a study found that American soldiers who were angry "were more likely than those who [were] not to kick and hit non-combatants, verbally abuse noncombatants, and unnecessarily damage property."[231]

5. Social Biases

Because soldiering is a group activity, all of the individual biases discussed above are likely to be compounded by what cognitive psychologists call "groupthink": the tendency of group members to seek unanimity, even if the group's unanimous position will not reflect their individual beliefs.[232] Groupthink not only leads members to fail to adequately consider alternative explanations and courses of action,[233] it generally causes "a deterioration of mental efficiency, reality testing, and moral judgement."[234] According to Barbara Adams et al., groupthink is strongest in groups that suffer "high stress from external threats" and are "highly cohesive, have strong directive leadership, are under time pressure, and have an important complex decision to make"—a description that reads like it was written specifically for the military.[235]

6. Debiasing

Although critics of autonomous weapons sometimes acknowledge the factors that undermine rational decision-making by soldiers,[236] they rarely take them seriously. Most simply ritually assert that, whatever flaws humans might have, machines will never possess the kind of human qualities IHL compliance

[228] Stanley & Larsen, *supra* note 198, at 78–79.

[229] *See* Mosier & Fischer, *supra* note 216, at 244.

[230] Lerner et al., *supra* note 219, at 803.

[231] Deanna L. Messervey et al., *Making Moral Decisions Under Stress: A Revised Model for Defence*, 21 CANADIAN MIL. J. 38, 41 (2021).

[232] *See* ADAMS ET AL., *supra* note 173, at 39.

[233] *See id.*

[234] IRVING L. JANIS, GROUPTHINK: PSYCHOLOGICAL STUDIES OF POLICY DECISIONS AND FIASCOES 9 (2nd ed., 1982).

[235] ADAMS ET AL., *supra* note 173, at 39.

[236] *See, e.g.*, Heyns, *supra* note 1, at 10 ("LARs will not be susceptible to some of the human shortcomings that may undermine the protection of life. Typically they would not act out of revenge, panic, anger, spite, prejudice or fear.").

requires.[237] Sharkey is one of the few exceptions. In his view, the appropriate response is to train human soldiers to think better, not to develop better killer robots: "[r]ather than funding technological 'hopeware', we need to direct funding into finding out where and when warfighters' ethical reasoning falls down and provide significantly better ethical training and better monitoring and make them more responsible and accountable for their actions."[238]

Unfortunately, the available evidence suggests that it is almost impossible to meaningfully debias the judgment of soldiers, particularly once they are in combat. The U.S. Army, for example, has devoted significant resources to trying to improve decision-making.[239] Its conclusion about a number of debiasing techniques, such as teaching soldiers to "consider the opposite of whatever decision they are about to make" or "take an outsider's perspective" on a possible decision,[240] is sobering:

> [T]he fundamental notion behind these approaches is, in some fashion, to subsume intuition with analysis. This comes into direct conflict with the nature of many of the decision making scenarios that military professionals in the future operating environment expect to face. The complexity and uncertainty of these situations may not afford individuals the luxury of time or collaboration that the methods detailed above require, particularly for individuals operating at the tactical level where commanders and soldiers must rely on quick, often inherently intuitive, decisions.[241]

Canada's Department of Defence is similarly pessimistic, noting that "whole scale [sic] remediation efforts might be ineffective"[242] because "it is clear that 'broad stroke' solutions to these systematic errors and biases will not be possible, as they are embedded within individual human psychology and will not be easily remediated."[243]

[237] *See, e.g.*, Asaro, *supra* note 129, at 700 ("The very nature of IHL, which was designed to govern the conduct of humans and human organizations in armed conflict, presupposes that combatants will be human agents. It is in this sense anthropocentric."); HUMAN RIGHTS WATCH, *supra* note 88, at 34 (arguing that "even if a robot could reach the required level of reason, it would fail to have other characteristics—such as the ability to understand humans and the ability to show mercy—that are necessary to make wise legal and ethical choices beyond the proportionality test").

[238] Noel E. Sharkey, *The Evitability of Autonomous Robot Warfare*, 94 INT. REV. RED CROSS 787, 796 (2012).

[239] *See* RODMAN, *supra* note 152, at 24 ("The US Army has expended considerable energy elaborating on decision making, detailing how and why it is important, and exploring tools, methods and initiatives to improve decision making among its personnel.").

[240] *Id.* at 21.

[241] *Id.* at 22.

[242] ADAMS ET AL., *supra* note 173, at 136.

[243] *Id.* at 137.

These gloomy assessments are consistent with a significant amount of debiasing research. In terms of statistical reasoning, for example, meta-analyses indicate that "simply encouraging higher levels of attention and care in making decisions" has no appreciable effect on the many biases associated with the three main heuristics: availability, representativeness, and affect (emotions).[244] Educating people about the existence of cognitive biases[245] and providing feedback on decision-making[246] are similarly unhelpful. Training people in statistical reasoning has had some success for simple tasks,[247] but it "has not typically been fully tested in complex environments using unfamiliar and abstract rules"[248]—a perfect description of IHL-compliant targeting. And although informing people that they will be held accountable for their actions can improve performance to some extent, accountability "does not wholly eliminate biased responding."[249]

Efforts to ameliorate the negative effect emotion has on decision-making have fared little better. Asking soldiers to suppress emotion while making decisions not only doesn't work,[250] it comes with significant costs, such as decreased ability to recall information, and can "even escalate the negative emotions it was intended to suppress."[251] Equally useless are "[s]aturating the decision maker with cognitive facts about a particular decision domain"[252] and encouraging emotional self-awareness.[253] These empirical findings are not surprising, because the "mental contamination" that leads to emotion affecting judgment "arises because of mental processing that is unconscious or uncontrollable."[254] Indeed, that is precisely why the only promising debiasing technique for emotion is "altering the structure of the choice context," such as organizing a cafeteria to ensure healthy options appear before junk food.[255] Militaries, however, do not choose the choice context their soldiers encounter in combat. The context chooses them.

Even if debiasing techniques were more successful, they would still have to grapple with another inherent aspect of combat: namely, that it involves decisions so complex that they are likely to be distorted by multiple cognitive biases. One example of this, the attack on the MSF hospital that resulted from both base-rate bias and confirmation bias, was mentioned earlier. An even more striking example

[244] *Id.* at 55.
[245] *Id.* at 116 (noting that "attempts to raise awareness about biases have not been shown to be consistently effective").
[246] *Id.* at 116–17 (noting that "the use of feedback to improve confidence calibration in individuals [is] less likely to be an effective strategy within a general domain").
[247] *Id.* at 121 ("Training has been shown to be effective with some forms of bias, but unfortunately, its effects have primarily been tested with relatively simple tasks.").
[248] *Id.* at 120.
[249] *Id.* at 118.
[250] *See* Lerner et al., *supra* note 219, at 812.
[251] Stanley & Larsen, *supra* note 198, at 84.
[252] Lerner et al., *supra* note 219, at 813.
[253] *Id.* (noting that "even when people are motivated, attaining accurate awareness of their decision processes is a difficult task").
[254] *Id.*
[255] *Id.* at 814.

comes from a study involving trained tactical action officers in a shipboard anti-air-warfare (AAW) unit. The AAW officers were asked to monitor incoming aircraft in six 18-minute test scenarios and determine what level of threat the aircraft posed. Each scenario involved multiple types of information: radar emissions; correspondence to commercial flight lanes and commercial flight schedules; intercepted verbal communications; responses to warnings on civil and military radio; and changes in altitude and airspeed. The information differed in each scenario in order to test the presence of various cognitive biases.[256] The results of the study were discouraging, to say the least:

> [E]xperienced subjects, making decisions in a realistic simulation of a ship-board environment, were subject to several of the biases observed in more artificial settings. AAW decision makers select hypotheses based on representativeness of the evidence and availability of the hypotheses, while ignoring base rate and not taking the reliability of information sources into account. Order of evidence effects were particularly strong in this test. The subjects also demonstrated a clear confirmation bias, placing more emphasis and recalling more information that supported their final conclusions.[257]

Importantly, differences in experience and training had almost no effect on the performance of the AAW officers. "Experienced tactical officers were subject to the same biases as the novice trainees."[258]

In short, critics of autonomous weapons cannot compensate for the irrationality of human decision-making by arguing, as Sharkey does, that soldiers should just be trained better. There is no evident way to eliminate or even substantially reduce the cognitive biases most relevant to soldiers, much less debias multiple ones at the same time.

D. Meaningful Human Control vs. Meaningful Human Certification

Despite these well-documented limits on human decision-making, critics routinely argue that, if they are to be used in combat, autonomous weapons must always be subject to "meaningful human control" (MHC).[259] Such control,

[256] *See* Bruce M. Perrin, Barbara J. Barnett & Larry C. Walrath, *Decision Making Bias in Complex Task Environments*, 37 PROC. HUM. FACTORS & ERGONOMICS SOC'Y ANN. MEETING 1117, 1118 (1993).

[257] *Id.* at 1120.

[258] *Id.*

[259] *See, e.g.*, Parsley, *supra* note 148, at 434 ("Introduced by the NGO Article 36 in 2014, MHC has emerged as the site of an initial regulatory consensus, expressing the 'widespread understanding that both the legal and ethical acceptability of a weapon system would require some kind of human control.'").

however, creates more problem than it solves. What is needed instead is meaningful human certification.

1. Meaningful Human Control

Critics of autonomous weapons acknowledge that little consensus exists over how meaningful human control should be defined.[260] Nevertheless, most proposals follow the three conditions identified by the International Committee for Robot Arms Control (ICRAC): (1) a human must have "full contextual and situational awareness of the target area and be able to perceive and react to any change or unanticipated situations"; (2) a human must have "sufficient time for deliberation on the nature of the target," including determining military necessity and proportionality; and (3) a human must have "a means for the rapid suspension or abortion of the attack."[261]

A thorough analysis of meaningful human control is beyond the scope of this article, but it is also unnecessary. As the three ICRAC requirements indicate, the human who exercises MHC is no less expected to be a paragon of rationality than a soldier engaged in physical combat.[262] And the ability of humans to fulfill that expectation is no less assumed.[263]

Nearly all of the cognitive limits on the performance and decision-making of human soldiers discussed above apply equally to the human who is expected to exercise meaningful control over an autonomous weapon during combat. Those limits alone are sufficient to call into question the idea that MHC is a solution to the "problem" of AWS. But the case against MHC does not end there. On the contrary, there are at least three cognitive biases unique to human/machine teaming that make the case even stronger.

[260] *See, e.g.*, Daniele Amoroso & Guglielmo Tamburrini, *In Search of the "Human Element": International Debates on Regulating Autonomous Weapons Systems*, 56 INT'L SPECTATOR 21 (2021) ("[I]t is far from clear, even among those favouring an MHC requirement, exactly what its actual content should be.").

[261] Frank Sauer, *ICRAC Statement on Technical Issues to the 2014 UN CCW Expert Meeting*, ICRAC INT'L COMM. ROBOT ARMS CONTROL (May 14, 2014), icrac.net/2014/05/icrac-statement-on-technical-issues-to-the-un-ccw-expert-meeting [https://perma.cc/WE4U-2ZN2].

[262] *See, e.g.*, Williams, *supra* note 148, at 6 ("MHC as a key component of managing LAWS risks therefore privileges humans best able to navigate this complex, dangerous, and multi-faceted strategic environment. They will have to be astute strategists; skilled diplomats; fully versed in military doctrine, operations, and tactics; and calm calculators of utility maximization able to balance the dilemmas, even trilemmas in pressured situations.").

[263] *See, e.g.*, Parsley, *supra* note 148, at 436 ("The human, in MHC, is entirely presupposed. Though remarkable, this omission is commonplace. The presupposition that our interlocutors (or even ourselves) are human is fundamental—and perhaps humanness even consists in simply not ruining this illusion.").

The first is automation complacency, also known as overtrust:

> Studies have consistently shown that there is a tendency for humans to place uncritical trust in computer-based decision systems (automation bias), as we have a tendency to ignore, or not search for, contradictory information in light of a computer-generated solution, especially in "time-critical decision support systems." This applies not only to autonomous or automated systems, but also to 'mixed-mode' systems where the human is in the loop to review the decisions, and is particularly pronounced in systems with high levels of autonomy in decision making, such as AI systems. In other words, the cognitive asymmetry between humans and AI systems produces outcomes that are invariably skewed in favour of machine decisions.[264]

Autonomous weapons are tailor-made to engender automation complacency, given their high level of autonomy and use in "time-critical" situations such as combat. Human operators are thus unlikely to challenge AWS targeting decisions even when they are incorrect. That would not be so concerning if automation complacency could be debiased, but that is not the case: studies have shown that such complacency "is actually pernicious, and perhaps even intractable,"[265] because it is not affected by an operator's expertise,[266] affects both individuals and teams,[267] and cannot be prevented by warning operators to verify machine decisions because they might be incorrect.[268]

Reminding operators of the fallibility of an automated system can, in fact, lead to the second human/machine cognitive bias: undertrust.[269] When a human operator is unduly skeptical of an automated system's reliability, he is likely to ignore relevant information it provides and to override its decisions— with potentially catastrophic consequences. Indeed, scholars have invoked undertrust to explain both the *USS Vincennes*' downing of Iranian Air Flight 655 in 1998[270] and the *USS John S. McCain*'s collision with a Liberian tanker in 2017. In the former

[264] Elke Schwarz, *Autonomous Weapons Systems, Artificial Intelligence, and the Problem of Meaningful Human Control*, 5 PHIL. J. CONFLICT & VIOLENCE 64 (2021); *see also* HUMAN RIGHTS WATCH, *supra* note 88, at 12–13 ("During the actual operation of the machine, the operator really only exercises veto power, and a decision to override a robot's decision must be made in only half a second, with few willing to challenge what they view as the better judgment of the machine.").

[265] John Zerilli et al., *Algorithmic Decision-Making and the Control Problem*, 29 MINDS & MACHINES 555, 556 (2019).

[266] *See* Raja Parasuraman & Dietrich H. Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, 52 HUM. FACTORS 381, 397 (2010).

[267] *Id.*

[268] *Id.*

[269] VINCENT BOULANIN ET AL., LIMITS ON AUTONOMY IN WEAPON SYSTEMS: IDENTIFYING PRACTICAL ELEMENTS OF HUMAN CONTROL 19 (2020) (noting that "[u]nder-trust is the opposite: it is the propensity for human operators to place insufficient reliance on an autonomous system").

[270] *Id.*

case, operators ignored information from the Aegis Combat System that the airplane was ascending, not descending;[271] in the latter, the captain refused to trust a new automated navigation system because it had proven "glitchy."[272]

The third and final human/machine cognitive bias concerns the attention of a human operator. Constantly monitoring a battlefield via "screens, sensors, scopes" necessarily increases an operator's cognitive load, depleting executive functioning and increasing "vulnerability to emotion dysregulation."[273] Ideally, therefore, human/machine teaming would require the operator to exhibit maximum attention only at particularly critical moments,[274] such as when an autonomous weapon selects a target to attack. But that is obviously infeasible in combat, because an operator cannot know precisely when an attack will occur. The operator of an AWS thus has two choices: maintain constant vigilance, accepting the cognitive overload that comes with it; or try to react when an AWS selects a target. The former is impractical,[275] and the latter is counterproductive, as humans are extremely likely to make errors when they need to quickly switch from a low-cognitive-effort task to a high-cognitive-effort one.[276]

These teaming issues will only get worse as AWS technology improves. Automation complacency, for example, is positively correlated with accuracy: the more reliable a machine proves to be, the more likely its operator will uncritically accept the decisions it makes.[277] Similarly, as autonomous weapons become increasingly involved in combat and their speed, geographic range, and loitering ability increases, the attentional demands that meaningful human control places on human operators will become ever greater. Speed is likely to be a particularly acute problem: even critics of AWS acknowledge that the tempo of combat with machines will eventually become so fast—what the U.S. calls "hyperwar" and

[271] Anthony Tingle, The Human-Machine Team Failed Vincennes, U.S. NAVAL INST. (July 2018), https://www.usni.org/magazines/proceedings/2018/july/human-machine-team-failed-vincennes [https://perma.cc/W5VP-4FXL].

[272] Rebecca Crootof, *AI and the Actual IHL Accountability Gap*, CTR. FOR INT'L GOVERNANCE INNOVATION (Nov. 28, 2022), https://www.cigionline.org/articles/ai-and-the-actual-ihl-accountability-gap/ [https://perma.cc/N28R-CTYQ].

[273] Stanley & Larsen, *supra* note 198, at 89.

[274] *See* Zerilli et al., *supra* note 265, at 565 ("Ideally, only those parts of a decision should be automated that leave the human operator with something vital and absorbing to do.").

[275] *See, e.g.*, BOULANIN ET AL., *supra* note 269, at 19 (noting that "[s]imply monitoring systems holds people's attention poorly").

[276] *See* Zerilli et al., *supra* note 265, at 565 (noting that "effective intervention can be enormously difficult when the operator has to shift from low to high level cognitive effort within a very short window").

[277] *See, e.g.*, David Lyell & Enrico Coiera, *Automation Bias and Verification Complexity: A Systematic Review*, 24 J. AM. MED. INFORMATICS ASS'N. 423, 424 (2017) ("Interestingly, high levels of system accuracy may inadvertently contribute to AB. This may be because accuracy engenders trust, and it has been shown that users who have greater trust in automation are less likely to detect automation failures.").

China calls "battlefield singularity"[278] —that humans will no longer be able to effectively monitor, much less control, AWS target selection.[279]

Swarming is also likely to reveal the futility of meaningful human control. Numerous states are currently developing groups of "small, low-cost munitions" that "can share target information and autonomously coordinate their strikes after launch,"[280] and at least two—the U.S. and Israel—already have the technological ability to deploy swarms in combat.[281] Humans will struggle to control the targeting selection of *one* AWS, for the reasons discussed above. Controlling swarms of them will be cognitively impossible.[282] That lack of control will almost certainly lead to automation complacency, because studies have shown that complacency is "typically found under conditions of multiple-task load, when manual tasks compete with the automated task for the operator's attention."[283]

In short, the idea that autonomous weapons should always be subject to meaningful human control is problematic. Such control will be undermined not only by the same cognitive biases that distort the decision-making of soldiers who participate directly in combat, but also by a number of biases that are unique to human/machine teaming. And if those issues are not themselves sufficient to reject MHC, inevitable developments in AWS technology, particularly speed and swarming, complete the case against it. Simply put, modern warfare involving killer robots will soon become so fast and so complicated that it will exceed not only "meaningful" human control, but human control itself. That pragmatic truth has to be factored into the case against autonomous weapons.

[278] CHRISTIAN RUHL, AUTONOMOUS WEAPON SYSTEMS & MILITARY AI 14 (2022).

[279] *See, e.g.*, Schwarz, *supra* note 264, at 66 (noting that, in the future, the human operator "may not even have enough time to process the information required to exercise control. Indeed, the vastly divergent times scales in human and machine operations might even rule out effective control of our machines"); Robert Sparrow, *Killer Robots*, 24 J. APPLIED PHIL. 68 (2007) ("[A]s AI technology improves, a human operator may prove not merely redundant but positively disadvantageous in such systems . . . It seems likely that sometime in the not-too-distant future, the time available to make survival critical decisions will often be less than the time required for a human being to make them."); BOULANIN ET AL., *supra* note 269, at 22 ("[I]t might be preferable to keep humans in direct (remote) control of the targeting functions—though that may not be possible in some narrow operational situations where human reaction speed is a limiting factor.").

[280] WORK, *supra* note 34, at 8.

[281] *See* Matilda Arvidsson, *The Swarm That We Already Are: Artificially Intelligent (AI) Swarming "Insect Drones", Targeting and International Humanitarian Law in a Posthuman Ecology*, 11 J. HUM. RTS. & ENV'T 11, 115 (2020) ("[S]warming AI drones are currently developed by prolific powers around the globe, designed for military and surveillance purposes.").

[282] *See, e.g.*, Sharkey, *supra* note 86, at 378 ("[H]uman decision making will be too slow and not able to react to the control of several aircraft at once. With the increasing pace of the action and with the potential of several aircraft to choose targets at the same time, it will not be possible to have the human make all of the decisions to kill.").

[283] Parasuraman & Manzey, *supra* note 266, at 390.

2. Meaningful Human Certification

Many critics are aware that meaningful human control is not a panacea for the problems ostensibly created by autonomous weapons. Instead of questioning whether such control is desirable, however, they simply insist that states prohibit the use of AWS that cannot be meaningfully controlled. Thus Ingivild Bode and Tom Watts argue, based on their insightful study of automated and autonomous air-defense systems, that the "requisite conditions" of meaningful human control "should be codified in international law" because "they represent a technological Rubicon which should not be crossed as going beyond these limits makes human control meaningless."[284]

This response to the problem of meaningful human control makes sense only if meaningless human control will lead to unnecessary death and destruction. But that is not necessarily the case: given the irrationality of human decision-making, particularly in combat, warfare that eliminates human control—that is fought autonomously—promises *less* unnecessary death and destruction, not *more*. Put more simply: in terms of IHL compliance, humans are the problem, not the solution.

There is, of course, an obvious response to this argument: namely, that autonomous weapons do not completely remove humans from the loop, because humans are still responsible for their programming and activation. After all, this article has repeatedly argued that AWS are nothing more than highly sophisticated tools for giving effect to the intentions of their human operators. Are those operators—the programmers and activators—not subject to the same cognitive, physiological, situational, and emotional limits as human soldiers and humans asked to exercise meaningful human control?

To some extent, all humans are subject to the limits discussed above. There is, however, a critical difference between the humans who fight or exercise meaningful control and the humans who program and activate autonomous weapons: namely, their temporal location in the kill chain. Humans in the first category, soldiers and controllers alike, make decisions concerning the use of lethal force in the heat of battle. Humans in the second category, by contrast, make those decisions in the cooler moments before combat begins. In general, therefore, the decision-making of humans who program and operate AWS will be less distorted by cognitive biases, physiological limits, situational constraints, negative emotions, and human/machine teaming issues than the humans who fight or control them.

[284] INGIVILD BODE & TOM WATTS, MEANING-LESS HUMAN CONTROL: LESSONS FROM AIR DEFENCE SYSTEMS ON MEANINGFUL HUMAN CONTROL FOR THE DEBATE ON AWS 4–5 (2021).

a. Cognitive Biases

As discussed above, most cognitive biases are caused by the "intuitive and automatic, sub-conscious, associative, affective and heuristic-based" nature of Type 1 thinking.[285] Type 2 thinking, by contrast, is "slower, deliberate, rule-based and takes place under conscious control," which makes it much less likely to lead to mistaken decisions.[286] In fact, the only proven method for debiasing Type 1 thinking is to deliberately shift people to Type 2, taking advantage of Type 2's slower and more deliberate kind of thinking.[287]

The ability to engage in Type 2 thinking is particularly important in the military context, because, as the U.S. Army has noted, it encourages soldiers to "consider multiple options, debate with others, contemplate alternative perspectives, and come to logical and, ideally, thorough and effective conclusions."[288] Type 2 thinking, however, is essentially impossible during combat, where life-and-death decisions must be made quickly in environments marked by unpredictability, complexity, and uncertainty. Both human soldiers and humans exercising real-time control over targeting by autonomous weapons—what the concept of meaningful human control requires—will thus almost always rely on bias-laden Type 1 thinking when making targeting decisions.

Humans who program and activate autonomous weapons are in a very different situation. The lawful use of an AWS for offensive purposes involves "two layers of target identification": (1) selecting an IHL-compliant target, group of targets, or general class of target; and (2) ensuring that an AWS is technologically capable of completing the targeting mission at least as well as human soldiers.[289] Both of those decisions are made prior to the actual targeting itself, which means that they are far more likely to be the product of analytic Type 2 thinking than the real-time targeting decisions made by human soldiers or human AWS controllers. Unlike those individuals, the programmers and activators of autonomous weapons will normally have the time and resources—cognitive, material, informational—that slow and deliberate reasoning requires. They are thus far more likely to make accurate decisions.

b. Physiological Limits

Humans who program and activate autonomous weapons are also much less likely than human soldiers and humans who control AWS to encounter the physiological limits that undermine rational decision-making. To begin with, because programming and activation take place before a battle begins, the

[285] VAN DEN BOSCH & BRONKHORST, *supra* note 149, at 2.

[286] Croskerry, Singhal & Mamede, *supra* note 151, at ii60.

[287] *See id.* at ii61.

[288] RODMAN, *supra* note 152, at 13.

[289] *See* M.L. Cummings, *Lethal Autonomous Weapons: Meaningful Human Control or Meaningful Human Certification?*, IEEE TECH. AND SOC'Y MAG., Dec. 2019, at 24–25.

individuals who make the relevant decisions are unlikely to be as physically exhausted as soldiers who engage in close-up combat. Moreover, although exercising meaningful control over an AWS is also less physically demanding than engaging in actual combat, a controller is still much more likely to experience cognitive overload than a programmer or activator. Programming and activation are complex processes, but they do not involve the multitasking and attentional vigilance that meaningful human control requires. Indeed, there is little cognitive difference between a human soldier deciding whether to fire at a target and a human deciding whether to permit an autonomous weapon to fire. Both require—to recall the ICRAC definition of MHC—"full contextual and situational awareness of the target area," as well as the ability "to perceive and react to any change or unanticipated situations."

c. Situational Constraints

Similar considerations apply to the situational constraints that distort rational decision-making. Unlike human soldiers, AWS programmers and activators do not have to make decisions in the noise and physical heat of combat. And unlike both human soldiers and humans who control autonomous weapons, AWS programmers and operators do not have to deal with the same kind of time-pressures, such as the need to make (increasingly) quick decisions about whether to fire on a particular target.

d. Negative Emotions

All soldiers are likely to feel both stress and anger while participating in conflict. But that does not mean the humans who program and activate an autonomous weapon will normally feel the same amount of stress and anger as humans who participate in combat (close-up or remote) or who oversee an AWS's use of lethal force. On the contrary, the operators are likely to feel less of each emotion simply by virtue of being away from the battlefield (unlike a human soldier) and by not having to make decisions in real-time concerning whether to take human life (unlike a human who ostensibly exercises meaningful control over an autonomous weapon). The decision-making of programmers and activators is thus far more likely to involve Type 2 thinking than Type 1.

e. Teaming Issues

Finally, human/machine teaming issues are almost certainly less acute for humans who program and activate autonomous weapons than for humans who control them. Although both categories might overtrust or undertrust an AWS, programmers and activators will not have to make time-sensitive decisions about whether to permit the use of lethal force. Instead, they will be able to take whatever time is necessary to determine whether a previously used autonomous weapon functioned reliably enough to be tasked with future missions. That luxury of time, which humans trying to meaningfully control an AWS necessarily lack, will reduce (though certainly not eliminate) the likelihood of programmers and activators either uncritically assuming that an AWS is reliable or uncritically refusing to trust that an AWS will properly carry out their instructions. Moreover, because programmers and activators carry out their responsibilities before an autonomous weapon engages in targeting, they will not have to worry about the difficulty of controlling extremely fast AWS or AWS that act as swarms.

f. Certification Not Control

M.L. Cummings describes the *ex ante* process of selecting an IHL-compliant target and ensuring that an autonomous weapons is capable of completing the targeting mission as "meaningful human certification."[290] In her view, the future of effective yet humane warfighting lies precisely in guaranteeing such certification, "not insisting on an illusory concept of meaningful human control."[291] For all the reasons discussed above, Cummings' conclusion is sound: the process of certifying an autonomous weapon is far less likely to be distorted by cognitive biases, physiological limits, situational constraints, negative emotions, and human/machine teaming issues than the process of trying to control one. Meaningful human control *accentuates* those problems; it does not solve them.

III. The Necessity of Human Compassion

One of the most compelling arguments in favor of autonomous weapons is that they cannot feel the negative emotions that—as the previous section explained—undermine decision-making in combat.[292] Even critics who want to ban killer robots acknowledge this advantage. Heyns, for example, notes that autonomous weapons "would not act out of revenge, panic, anger, spite, prejudice or fear" and, "unless specifically programmed to do so . . . would not cause intentional suffering on civilian populations, for example through torture."[293]

That said, robots also do not feel *positive* emotions. That lack underlies an extremely common consequentialist critique of autonomous weapons: namely, that soldiers must be human because the ability to act with compassion is necessary to ensure that conflict remains as humane as possible. Sometimes this argument is legal, such as when Asaro writes that IHL "explicitly requires combatants . . . to apply compassion and judgement in an explicit appeal to their humanity."[294] The legal version of the compassion argument, however, is clearly incorrect. The argument can only be directed at killing combatants, because it is always unlawful

[290] *Id.* at 26.
[291] *Id.* at 24.
[292] *See* Ronald C. Arkin, Governing Lethal Behavior in Autonomous Robots 6 (2009) (noting that AWS "can be designed without emotions that cloud their judgment or result in anger and frustration with ongoing battlefield events").
[293] Heyns, *supra* note 1, at 10.
[294] Asaro, *supra* note 129, at 700; *see also* Guarini & Bello, *supra* note 89, at 137–38 (claiming that "the laws of war require compassion").

to intentionally kill civilians. But nothing in conventional or customary IHL *requires* sparing a combatant who is otherwise lawfully targetable. Unless he is *hors de combat*—a status determined by IHL itself[295]—a combatant can be killed anywhere at any time. So although a soldier may choose not to kill a combatant even when he is entitled to do so, that is an ethical decision, not a legal one. Even many critics of autonomous weapons admit as much.[296]

The ethical argument for compassion is far more common than the legal one.[297] It comes in two forms, one that focuses on civilians and one that focuses on combatants. Human Rights Watch provides an example of the first version when it claims that "although fully autonomous weapons would not be swayed by fear or anger, they would lack compassion, a key safeguard against the killing of civilians."[298] Leveringhaus provides an example of the second when he argues that because "the enemy about to be targeted is still a fellow human being with one life to live . . . retaining human agency at the point of force delivery, thereby protecting the freedom not to pull the trigger, push the button, or throw a grenade, is essential for retaining our humanity in exactly the situation that challenges it the most: war."[299]

Neither version of the compassion argument is compelling. The first problem, which applies to both, is one we have seen before: anachronism. The quote from Human Rights Watch is predicated on close-up combat, where a soldier

[295] *See* AP I, *supra* note 85, art. 41.

[296] *See, e.g.*, Leveringhaus, *supra* note 76, at 349 (noting that "not shooting at a legitimate enemy target may, under certain circumstances, be morally desirable but it is not obligatory"); GEISS, *supra* note 78, at 16 ("The problem of deliberation on the basis of moral-legal fundamental principles may therefore be exaggerated: it is not a matter of arriving at one's own judgement based on one's own deliberations. On the contrary, soldiers are only supposed to apply those rules that the international community has established on the basis of universally valid considerations.").

[297] *See, e.g.,* Heyns quoted in Birnbacher, *supra* note 58, at 12 (arguing that the "most offensive part" of being killed by a machine is not death itself, "but rather the deprivation of hope for some kind of mercy or reprieve that this technology brings. Since there is no deliberative process, there is no possibility of a higher appeal, no prospect of human empathy"); Amoroso & Tamburrini, Ethical, *supra* note 15, at 8 ("The ensuing death-or-life decision could hardly be overridden when the AWS is about to actually release force, with the consequence that the human target would be somehow 'written off' without the (even slightest) hope of changing his/her fate."); Korać, *supra* note 19, at 51 ("Emotions and empathy, as drivers of prosocial behaviour and moral sensitivity, are a major obstacle to killing in war."); GEISS, *supra* note 78, at 18 ("[A] person attacked by an autonomous weapons system basically lacks the opportunity to appeal to the attacker's humanity. Factors such as dignity or empathy are removed from the equation.").

[298] HUMAN RIGHTS WATCH, *supra* note 21, at 7; *see also* Denise Garcia, *Killer Robots: Why the US Should Lead the Ban*, 6 GLOB. POL'Y 57, 59 (2015) ("Human emotion remains one of the best safeguards against the killing of civilians and is a central constraint on barbarity.").

[299] Leveringhaus, *supra* note 76, at 350; *see also* ICRC, POSITION ON AUTONOMOUS WEAPONS SYSTEMS 8 (2021) (arguing that "in decisions about life and death," the use of autonomous weapons "removes the possibility for restraint, a human quality that means people may decide not to use force even if it would be lawful"); BOULANIN ET AL., *supra* note 269, at 13 (noting that compassion "enables those persons to exercise restraint or mercy, even when that course of action (restraint/mercy) is not strictly required for them to comply with the law").

can look into the eyes of the civilian or enemy soldier he is about to kill and stay his hand. Most modern warfare, however, is remote. Compassion is irrelevant to the bombardier dropping bombs from 25,000 feet or the soldier firing a HIMARS rocket at a target 50 miles away, and it is scarcely more relevant to the UAV operator manipulating a joystick on a different continent, whose human targets appear as little more than moving smudges on a video screen.

Most wars still involve infantry, of course, so there may well be situations in which the human capacity for compassion means the difference between a civilian's life and death. But the importance of that possibility should not be overstated. To begin with, compassion cannot prevent the accidental or mistaken killing of civilians, which is by far the most common cause of civilian death during combat. Gregory McNeal estimates, for example, that 70% of civilian deaths in Afghanistan and Iraq resulted from misidentification.[300] Moreover, insofar as the possibility of compassion is offered as an argument against potentially more discriminating autonomous weapons, the calculus must take into account all of the situations in which a human soldier *fails* to exercise compassion and intentionally kills a civilian. That number is almost certainly far higher, given how common civilian massacres are in warfare. The Early Warning Project, for example, lists 20 conflict locations in the world where more than 1,000 civilians are currently being intentionally killed each year.[301]

To be fair, it is unlikely that AWS critics who advance the civilian compassion argument are imagining a situation in which a soldier who is hell-bent on killing a civilian suddenly has a change of heart. Instead, they are almost certainly thinking about the possibility of compassion leading a soldier to disobey an unlawful order to kill civilians—something an autonomous weapon ostensibly could not do. A few scholars, in fact, make this argument explicitly. Duncan Purves et al., for example, claim that it would be far easier to make AI carry out immoral or criminal orders than it is to get human soldiers to carry out such orders, because "[i]f an AWS cannot make moral judgments, [it] cannot resist an immoral order in the way that a human soldier might."[302] Similarly, Johnson and Axinn claim that, unlike human soldiers, "robots have no basis for making" a judgment that an order is unlawful because they lack "their own values."[303]

It is far from clear, however, that an autonomous weapon could not be programmed to disobey an illegal order. Although designing a truly "virtuous

[300] Gregory S. McNeal, *Targeted Killing and Accountability*, 102 GEORGETOWN L. J. 681, 738 (2014); *see also* LARRY C. LEWIS, REDEFINING HUMAN CONTROL: LESSONS FROM THE BATTLEFIELD FOR AUTONOMOUS WEAPONS 4 (2018) (estimating that 50% of civilian casualties in Afghanistan were caused by misidentification).

[301] *Ongoing Mass Killing*, EARLY WARNING PROJECT, https://earlywarningproject.ushmm.org/ongoing-mass-killing [https://perma.cc/PJG9-SK94] (last visited Nov. 2, 2023).

[302] Purves, Jenkins & Strawser, *supra* note 20, at 858.

[303] Johnson & Axinn, *supra* note 53, at 135.

robot" may never be possible,[304] a number of roboticists believe that, "as a proxy for a full-fledged morality,"[305] fundamental rules of IHL could be programmed into autonomous weapons in a way that prevents them from being overridden by an operator or commander.[306] Thus programmed, an AWS would be unable to attack any target that it determined was civilian, even if instructed to do—the machine equivalent of the ability to disobey an unlawful order.

If autonomous weapons could be programmed to never attack targets they determine are civilian—and it is important to acknowledge that, as discussed earlier, there will be some situations in which a machine will struggle with the principle of distinction—they would represent an *improvement* over human soldiers.[307] Human soldiers have the capacity to disobey illegal orders, but they also have the capacity to carry them out. And given the significant number of civilians who are deliberately murdered in armed conflict each year, soldiers seem far more likely to carry out illegal orders than disobey them.

The combatant-centered compassion argument is even less convincing. The issue here is not whether a soldier will kill a combatant who is surrendering, already captured, or wounded. Killing a combatant who is *hors de combat* is never lawful. The situation imagined by AWS critics is one in which, during combat, a soldier chooses not to kill an enemy combatant he has in his sights.[308] That reaction is certainly possible, because humans have an "innate resistance to killing."[309] Statistics nevertheless indicate that, as a result of their training and instinct for self-preservation, nearly 90% of soldiers will kill during combat when legally entitled to do so.[310] The need to maintain the mere possibility of compassionate non-killings

[304] *See, e.g.*, PATRICK LIN, GEORGE BEKEY & KEITH ABNEY, AUTONOMOUS MILITARY ROBOTICS: RISK, ETHICS, AND DESIGN 40 (2008) (noting that "many technological thresholds must be crossed before the development of a virtuous robot becomes a serious possibility").

[305] *Id.* at 42.

[306] *See, e.g.*, ARKIN, *supra* note 292, at 179 (noting that such rules "will be relegated to . . . long-term memory (LTM) for those constraints which persist over all missions" and that "[c]hanges in LTM, that encode the LOW, [will] require special two-key permission" to override); LIN, BEKEY & ABNEY, *supra* note 304, at 42 ("For military robots, that virtuous character will likely involve ensuring that the LOW and ROE are programmed in (which may differ from mission to mission) and steadfastly obeyed.").

[307] *Cf.* LIN, BEKEY & ABNEY, *supra* note 304, at 53 ("If this or any other ROE does violate the LOW, the ethical result of using robots may be a moral improvement, since robots properly programmed to never violate the LOW would refuse to follow immoral orders, unlike human soldiers who are trained to unfailingly follow all orders.").

[308] *See, e.g.*, Leveringhaus, *supra* note 76, at 350.

[309] HUMAN RIGHTS WATCH, HEED THE CALL: A MORAL AND LEGAL IMPERATIVE TO BAN KILLER ROBOTS (2018).

[310] Statistics from the Vietnam War indicate, for example, that 90% of soldiers shoot at the enemy during combat. Gert-Jan Lokhorst & Jeroen van den Hoven, *Responsibility for Military Robots*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 145, 147–48 (Patrick Lin, Keith Abney & George A. Bekey eds., 2012).

thus does not count as a persuasive argument against autonomous weapons.[311] As George R. Lucas says:

> [w]e neither want nor need our unmanned systems to "be ethical," let alone "more ethical" or "more humane" than human agents. We merely need them to be safe and reliable, to fulfill their programmable purposes without error or accident, and to have that programming designed to conform to relevant international law.[312]

It is also important to question the basic assumption of the combatant compassion argument, which is that it is ethically desirable for soldiers not to kill the enemy even when they have the legal right. That assumption is not self-evident. Michael Walzer, for example, discusses a situation that occurred at Anzio during World War II in which a group of British soldiers spotted a German soldier who was too busy enjoying the beautiful spring day to realize that he was exposed to enemy fire.[313] After debating whether to kill him, the soldiers decided to scare him away instead. Echoing the ICRC, Walzer cites the decision not to kill the "naked soldier" as a positive example of compassion during combat.[314]

The soldiers' superior, however, had a very different perspective on their decision. "Sergeant Chesteron didn't laugh. He said that we should have killed the fellow, since his friends would now be told precisely where our trenches were."[315] The Sergeant's response explains why, however cruel it might seem, IHL permits targeting combatants anytime, anywhere, and with any amount of force: because they have both the right and the ability to shoot back. A soldier might feel kinship with an enemy soldier enjoying lovely weather. He might feel guilty about killing him in such an obviously vulnerable state. But that enemy soldier has been trained to kill—and if he is left alive now, that is precisely what he will do later.

The consequentialist argument for showing compassion to enemy soldiers is thus more complicated than its defenders admit. Indeed, one of the scholars who endorses the argument, Leveringhaus, acknowledges that "it does not imply that soldiers should always show pity or mercy. It may be inappropriate, for instance, to show pity or mercy towards members of a genocidal militia about to commit a massacre."[316] But he insists—as an argument against the emotionless targeting of autonomous weapons—that scenarios such as the one at Anzio explain why "we must not lose our sense of humanity, pity, or mercy in war, no matter how justified

[311] *Cf.* Birnbacher, *supra* note 58, at 121 (asking rhetorically, "is an attack by a terror bomber less cruel only because the commander of the aeroplane might in principle be merciful whereas an autonomous system would not—if, in fact, the hope that this happens is as futile in the one case as in the other?").

[312] George R. Lucas, *Automated Warfare*, 25 STAN. L. & POL'Y REV. 317, 332 (2014).

[313] MICHAEL WALZER, JUST AND UNJUST WARS: A MORAL ARGUMENT WITH HISTORICAL ILLUSTRATIONS 140–41 (4th ed. 2006).

[314] *Id.* at 139–41.

[315] *Id.* at 141.

[316] Leveringhaus, *supra* note 76, at 354.

a particular order is. This is unlikely to undermine the effective functioning of the military."[317]

There are two problems with Leveringhaus's position. First, soldiers have no way of knowing what the "naked soldier" will do later if they show compassion toward him now. It is entirely possible the German soldier spared at Anzio later executed Allied prisoners of war or murdered Jews. Second, showing compassion to enemy combatants is "unlikely to undermine the effective functioning of the military" *only insofar as it remains exceptional*. If enough soldiers refuse to kill the enemy despite being legally entitled to do so, their side will lose the war. If the losing side is the aggressor, that would be a good outcome. But if the losing side is defending itself from aggression, quite the opposite is true.

IV. The Necessity of Human Accountability

Another consequentialist[318] objection to autonomous weapons worth considering, also focused on IHL, is that their use will create "accountability gaps": situations in which a war crime has been committed but no human can be held accountable for it. According to Human Rights Watch, "[g]aps in criminal accountability for fully autonomous weapons would exist under theories of both direct responsibility and indirect responsibility (also known as command responsibility)."[319] That is a problem, Sparrow insists, because "[i]t will be unethical to deploy autonomous systems involving sophisticated artificial intelligences in warfare unless someone can be held responsible for the decisions they make."[320]

Although AWS-created accountability gaps are indeed a concern, Human Rights Watch's argument is overstated. For a consequentialist objection to have merit, it is not enough to show that accountability gaps will exist if states use autonomous weapons. Those gaps must be larger than the ones that currently exist when states use only human soldiers. If equivalent accountability gaps will exist regardless of whether states use AWS or human soldiers, accountability does not provide a reason to prohibit autonomous weapons.

[317] *Id.* at 355.

[318] Some AWS critics view this issue as deontological. *See, e.g.*, Amoroso & Tamburrini, *supra* note 15, at 5 ("AWS are likely to determine an accountability gap. And the latter is hardly reconcilable with the agent-relative moral obligation of military commanders and operators to be accountable for their own actions."). The argument is primarily consequentialist, however, because the failure to punish those responsible for war crimes will make it more likely that soldiers will commit such crimes in the future.

[319] HUMAN RIGHTS WATCH, *supra* note 21, at 18.

[320] Sparrow, *supra* note 279, at 74–75; *see also* SIMON CHESTERMAN, WE, THE ROBOTS? REGULATING ARTIFICIAL INTELLIGENCE AND THE LIMITS OF THE LAW 104 (2021) ("The stronger argument for meaningful human control is not that humans will make better decisions but that humans can be held to account.").

A. Direct Responsibility

The accountability objection begins with an unobjectionable premise: human soldiers can be held accountable for committing war crimes while autonomous weapons cannot. Both sides of the "killer robot" debate generally accept this premise, for an obvious reason: AWS are weapons, not moral agents.[321] They cannot act with the *mens rea* necessary for a war crime,[322] and they cannot be criminally punished.[323]

Nearly all AWS critics also acknowledge that if they are weapons, not moral agents, humans can be held directly responsible for how AWS are used.[324] Specifically, criminal responsibility will exist whenever a human—programmer or operator—intentionally uses an autonomous weapon to commit a war crime (direct intent) or activates one despite knowing that a war crime is virtually certain to result (knowledge) or could result (recklessness).[325] In such cases, there is no accountability gap—just as there is no gap when a soldier commits a war crime with a non-autonomous weapon.

Recognizing this, critics have argued that there is, in fact, a relevant difference between autonomous weapons and non-autonomous weapons in this regard: namely, that it will often be difficult to determine *which human* is responsible for a war crime committed with an AWS—something that is rarely the case when a crime is committed with a precision-guided munition or a rifle.[326] Critics refer to this as the "many hands" problem:

> [T]he list of potentially responsible individuals is quite long, as it includes the software programmer, the military commander in

[321] *See, e.g.*, Sparrow, *supra* note 279, at 74 (arguing that "existing autonomous weapons systems remain analogous to other long-range weapons" and thus cannot be considered full moral agents); Russell Buchan & Nicholas Tsagourias, *Autonomous Cyber Weapons and Command Responsibility*, 96 INT'L L. STUD. 645, 670 (2020) (noting that AWS "cannot be held criminally responsible because they are not moral agents").

[322] Buchan and Tsagourias are among a small group of legal scholars who believe AWS can act with *mens rea*. *See* Buchan & Tsagourias, *supra* note 321, at 670–71.

[323] *See, e.g.*, HUMAN RIGHTS WATCH, *supra* note 21, at 19.

[324] *See, e.g.*, AMOROSO ET AL., *supra* note 18, at 28 ("As AWS obviously cannot be held responsible as direct perpetrators, responsibility for their actions should be traced back to some persons in the decision-making chain.").

[325] *See,* MARTA BO, LAURA BRUUN & VINCENT BOULANIN, RETAINING HUMAN RESPONSIBILITY IN THE DEVELOPMENT AND USE OF AUTONOMOUS WEAPON SYSTEMS: ON ACCOUNTABILITY FOR VIOLATIONS OF INTERNATIONAL HUMANITARIAN LAW INVOLVING AWS 31 (2022). The Rome Statute limits criminal responsibility for the modes of participation other than command responsibility to acts committed with intent and knowledge. Under customary international law, recklessness sometimes also suffices. *Id.*

[326] To be sure, this will not always be the case. It may be difficult, for example, to determine which soldier executed a prisoner of war if none of his colleagues will inculpate him. Normally, though, it will be relatively clear which soldier "pulled the trigger" on the non-autonomous weapon used to commit a war crime.

> charge of the operation, the military personnel that sent the AWS into action or those overseeing its operation, the individual(s) who conducted the weapons review, or political leaders, as well as the manufacturer of the AWS and the procurement official . . . To the extent that no one actually pushes the "fire" button, and hence assumes at least prima facie responsibility in case of wrongdoing, AWS technology will put those involved in their use in the position to "pass the buck" to others.[327]

This argument cannot be dismissed, but it is significantly overstated. To begin with, "many hands" are only a problem if they prevent *anyone* from being held responsible for the actions of an autonomous weapon. When an individual uses an AWS to intentionally commit a war crime, it is difficult to imagine that there will be no trace of his order or programming in the machine's system.[328] Proving that an individual knew it was virtually certain or was aware of the possibility that an AWS might commit a war crime would indeed be more difficult, because of the nature of autonomy. "Where there is inherent unpredictability in the functioning of a system it may not be possible to assess the probability of a certain action, and so determining risk becomes problematic."[329] That said, the difficulty of inferring knowledge or recklessness from circumstantial evidence is a general problem in criminal law, not one unique to autonomous weapons.[330] Moreover, given what might be called the "predictable unpredictability" of autonomous weapons, an operator could hardly plead ignorance if he continued to use an AWS that he knew had acted unpredictably before—especially in a way that had violated IHL.

In fact, in terms of non-volitional mental states like knowledge and recklessness, "many hands" might make accountability *easier* to establish. As long as even one human can be held responsible when an AWS commits the *actus reus* of a war crime, there is no accountability gap. So, although one or more of the individuals involved in the programming or operation of an autonomous weapon might not have the subjective awareness war crimes require, another might. Consider, for example, a situation in which an AWS "unpredictably" confuses a

[327] AMOROSO ET AL., *supra* note 18, at 28; *see also* HUMAN RIGHTS WATCH, *supra* note 21, at 20 ("Each party could try to shift blame to the others in an attempt to avoid responsibility. Therefore, proving which party was responsible for the orders that led to the targeting of civilians might prove difficult, even if a user did intentionally employ fully autonomous weapons to commit a crime."); DANIELE AMOROSO & GUGLIELMO TAMBURRINI, WHAT MAKES HUMAN CONTROL OVER WEAPONS SYSTEMS "MEANINGFUL"? 7 (2019) ("People in this list may cast their defence against responsibility charges and criminal prosecution in terms of their limited decision-making roles, as well as of the complexities of AWS systems and their unpredictable behaviour in the battlefield.").

[328] *See, e.g.*, BO, BRUUN & BOULANIN, *supra* note 325, at 47 ("While the potential advantages may not be unique to AWS as such, auditable algorithms, digital trails and logs are among the technical features associated with AWS that contain the potential to help inform an investigation."). Indeed, "[m]any states have explicitly taken steps towards [making] 'traceability in AI' a priority." *Id*. at 46.

[329] ICRC, *supra* note 30, at 15–16.

[330] *See, e.g.*, Rebecca Dresser, *Culpability and Other Minds,* 2 S. CAL. INTERDISC. L. J. 41, 48 (1993).

police uniform for a military uniform and kills a group of policemen. The military officer who activated the AWS might have had no idea the machine was unable to accurately distinguish police uniforms from military ones, while the programmer who designed the machine's software and tested its object recognition might have been fully aware of that limitation. The military officer could not be convicted of a war crime, while the programmer could (via aiding and abetting). The autonomous weapon's act thus does not lead to an accountability gap.

To be sure, critics who argue that autonomous weapons create direct-responsibility accountability gaps rarely rely solely on the "many hands" problem. Most focus on a more problematic kind of situation: where an autonomous weapon commits the *actus reus* of a war crime in such an unexpected or unpredictable manner that no human would have thought the act was possible, much less intended it to happen. As Neha Jain notes, these kinds of errors are inevitable with autonomous weapons simply by virtue of their autonomy:

> [T]he AWS is designed to apply non-deterministic decision making in an open and unstructured environment. Thus, not only will pre-set rules designed by the programmer not be capable of capturing the complete range of scenarios that the AWS will encounter, but the AWS will also be able to adapt the means and methods it uses to achieve programmer-designed ends. This will inevitably lead to uncertainty and error, which cannot be fully predicted or controlled by the programmer.[331]

When an autonomous weapon makes such an error, there will indeed be an accountability gap, because no human—not even an unknown one—can be held responsible for what the AWS has done.[332]

The inevitability of this kind of accountability gap, however, does not validate the accountability objection to autonomous weapons. And the reason is simple: *the same accountability gap exists when a human soldier accidentally or mistakenly commits the* actus reus *of a war crime*. As with the human who controls an AWS, such commission gives rise to criminal responsibility only if the soldier intended to commit the *actus reus* or was at least aware that the *actus reus* could

[331] Neha Jain, *Autonomous Weapons Systems: New Frameworks for Individual Responsibility*, *in* AUTONOMOUS WEAPONS SYSTEMS: LAW, ETHICS, POLICY 303, 313 (Nehal Bhuta et al. eds., 2016); *see also* AMOROSO ET AL., *supra* note 18, at 28 ("[T]he complexities of weapon autonomy and the resulting behavioral unpredictability in partially structured or unstructured warfare scenarios are likely to afford a powerful defense against criminal prosecution."); Benjamin Kastan, *Autonomous Weapons Systems: A Coming Legal "Singularity"?*, 2013 J. L. TECH. & POL'Y 45, 65 (2013) ("AWSs are complex new systems, which—despite the best efforts of designers, testers, and operators—will fail at one point or another.").

[332] BO, BRUUN & BOULANIN, *supra* note 325, at 34 ("Under the legal framework of individual criminal responsibility, an individual cannot be held responsible for the unpredictable behaviour and effects of an AWS because the ability to foresee is a necessary requirement of the mental elements of intent and knowledge.").

result from his actions. If he did not act with intent, knowledge, or recklessness, he cannot be convicted of a war crime—no matter how devastating the consequences of his actions.

This kind of accountability gap can occur in a variety of situations. Most obviously, accidents caused by technological failure are not limited to autonomous weapons, as even critics of AWS acknowledge.[333] During a protracted conflict, for example, it is likely that at some point a normally reliable cruise missile or precision-guided munition will malfunction and kill civilians. As long as the soldier who operated the weapon did not suspect that it might malfunction, he is not criminally responsible for the civilian deaths.

More quotidian accidents involving "dumb" weaponry lead to the same result. If a bombardier releases his payload over a military base fully expecting the bombs to strike only the base, he cannot be convicted of a war crime if wind unexpectedly pushes some of the bombs into a nearby civilian area.[334] Similarly, if a soldier fires a HIMARS missile at an ammunition depot believing that the attack will be proportionate because only a few civilians are present, he cannot be convicted of a war crime if dozens more civilians arrive at the depot while the missile is in flight.

Similar errors lead to non-criminal accidents in close-up combat. Consider a variation on the children-playing scenario discussed earlier: instead of playing with toy guns near the soldier, the children sneak up behind him and playfully jab them into his back. If the soldier spins around and reflexively pulls the trigger before he realizes he is not actually being attacked, he cannot be convicted of the war crime of murder even though, by killing the children, he committed its *actus reus*.[335]

Human soldiers also make factual mistakes that do not lead to criminal responsibility. If a soldier attacks a vehicle genuinely believing it to be an enemy tank, he cannot be convicted of a war crime if the vehicle turns out to be a Red Cross ambulance. The same is true of a UAV operator who kills a group of civilian mourners at a funeral because he genuinely believes they are Taliban fighters.[336]

[333] *See, e.g.*, *id.* at 14 ("Normal accident theory suggests that in tightly coupled complex systems—such as modern military weapon systems, including AWS—accidents are 'inevitable' over a long enough time horizon.").

[334] *See* Wood, *supra* note 106, at 3 (noting that "freak accidents resulting from weather, unforeseeable events, or weapons malfunctions which are outliers are all things for which responsibility need not be assigned").

[335] Individuals cannot be held criminally responsible for actions that are not voluntary, even when they bring about a consequence prohibited by a particular international crime. *See* KAI AMBOS, I TREATISE ON INTERNATIONAL CRIMINAL LAW: FOUNDATIONS AND GENERAL PART 274 (2013).

[336] *See* DOUGLAS GUILFOYLE, INTERNATIONAL CRIMINAL LAW 377 (2016) ("To take an easy example of a mistake of fact which negates the required mental element, consider the law of attacks on civilians or civilian objects as a war crime: the mental element required would be absent if the defendant had mistakenly identified the target as a military one.").

And the same is even true of a soldier who kills a small child carrying a stuffed animal because he mistakenly believes the stuffed animal is a grenade. In all three situations, the soldier who committed the *actus reus* of a war crime lacked the *mens rea* necessary for criminal responsibility because he did not act intentionally, knowingly, or recklessly. Indeed, because an honest mistake negates any subjective mental state,[337] none of the soldiers can be convicted of a war crime even if their beliefs—that the ambulance was a tank, that the car contained Taliban fighters, that the child was directly participating in hostilities—were grossly negligent.

In terms of direct responsibility, then, there is no difference between autonomous weapons and human soldiers in terms of accountability for war crimes committed by mistake or accident. The critical question is thus not whether AWS will make the kind of errors that lead to accountability gaps— they will—but whether AWS will make such unaccountable errors *more often* than human soldiers.[338] Some critics of autonomous weapons suggest as much,[339] but—as is too often the case with scholarship critical of killer robots —they focus solely on the limits of AWS technology, ignoring the cognitive limits on human performance and decision-making. Once those limits are taken into account, the direct-responsibility objection to autonomous weapons loses much of its force.

B. Command Responsibility

Critics also claim that the use of autonomous weapons will lead to accountability gaps in terms of command responsibility.[340] Those claims almost invariably turn on the idea that, because AWS are machines and not moral agents, human commanders could not be held criminally responsible for failing to prevent or punish war crimes AWS may commit. Amoroso et al., for example, say that "in

[337] *See, e.g.*, Otto Triffterer & Jens David Ohlin, *Article 32: Mistake of Fact or Law*, *in* ROME STATUTE OF THE INTERNATIONAL CRIMINAL COURT: A COMMENTARY 1161, 1170 (Otto Triffterer ed., 3rd ed. 2016) (noting that "a person who does not realize that the object is a hospital cannot be aware that the building is a protected target and therefore does not have the mental element required for committing a war crime under article 8(2)(b)(ix). In these cases, the question of whether criminal liability for negligence is at stake, does not affect the question of mistake of fact").

[338] Vincent C. Muller, *Autonomous Killer Robots Are Probably Good News*, *in* DRONES AND RESPONSIBILITY: LEGAL, PHILOSOPHICAL, AND SOCIOTECHNICAL PERSPECTIVES ON REMOTELY CONTROLLED WEAPONS 67, 76 (Ezio Di Nucci & Filippo Santoni de Sio eds., 2016) ("[I]f a technology produces rare cases of killings where no person is responsible, this do not by itself compel us to ban the use of this technology. A strong responsibility principle that allows no responsibility gaps at all is untenable in practice.").

[339] *See, e.g.*, BO, BRUUN & BOULANIN, *supra* note 325, at 14 ("It has been argued that AWS could be more prone to accidents, with sources of accidents arising from the risk of hacking, unexpected interactions with the environment, simple malfunctions and software errors.").

[340] *See, e.g.*, HUMAN RIGHTS WATCH, *supra* note 21, at 21 ("Significant obstacles would exist to establishing accountability for criminal acts committed by fully autonomous weapons under the doctrine of command responsibility."); AMOROSO ET AL., *supra* note 18, at vi (claiming that "it may be difficult to discern the conditions under which a commander's responsibility for a war crime in the use of AWS could arise").

case of misconduct by human soldiers, the commander may (and indeed must) exercise her or his punitive power over them—an option that is clearly precluded when the 'wrongdoer' is an AWS to which 'punishment' is a meaningless concept."[341]

Nearly all scholars, both critics and supporters of autonomous weapons, accept that command responsibility does not apply between human commanders and machine subordinates.[342] There is no question, however, that it does apply between human commanders and human subordinates who operate machines, autonomous and non-autonomous alike. As a result, if a human operator is responsible for a war crime "committed" by an AWS, normal principles of command responsibility will determine whether that soldier's commander will (also) be criminally responsible for it—just as they will when a soldier commits a war crime with a non-autonomous weapon.[343]

By contrast, command responsibility will not apply when there is an accountability gap at the level of direct responsibility. A commander can only be held responsible for failing to prevent or punish "crimes" committed by his subordinates,[344] and no crime has been committed when an autonomous weapon commits the *actus reus* of a war crime but none of the individuals involved in its programming and operation possess the requisite intent, knowledge, or recklessness. Critics are thus correct to argue that it will generally be impossible to hold military commanders responsible if, by virtue of its autonomy, an AWS accidentally commits the *actus reus* of a war crime.[345]

By itself, however, the absence of command responsibility in such situations does not validate the accountability objection to autonomous weapons. And once again the reason is simple: the same accountability gap exists when a human soldier accidentally or mistakenly commits the *actus reus* of a war crime. The soldier lacks the *mens rea* that war crimes require, so no war crime has been committed. And because no war crime has been committed, no commander can be held responsible for failing to prevent or punish it.

[341] AMOROSO ET AL., *supra* note 18, at 30.

[342] Once again, Buchan and Tsagourias are the exceptions. *See generally* Buchan & Tsagourias, *supra* note 321.

[343] *See, e.g.*, Schmitt, *supra* note 108, at 33 ("[T]he commander or civilian supervisor of that individual would be accountable for those war crimes if he or she knew or should have known that the autonomous weapon system had been so programmed and did nothing to stop its use, or later became aware that the system had been employed in a manner constituting a war crime and did nothing to hold the individuals concerned accountable.").

[344] *See* Rome Statute of the International Criminal Court, art. 28(a), July 17, 1998, UN Doc. A/CONF.183/9*, *reprinted in* 37 I.L.M. 999 (1998) ("A military commander or person effectively acting as a military commander shall be criminally responsible for crimes within the jurisdiction of the Court committed by forces under his or her effective command and control.").

[345] The rare exception being where its human operator was aware the machine might malfunction.

To validate the accountability objection, then, critics would have to show that, for some reason intrinsic to the nature of autonomous weapons, commanders are more likely to be held responsible for war crimes committed by human soldiers than for war crimes committed by human operators of autonomous weapons. Unless that is the case, the accountability objection applies to command responsibility generally—not to autonomous weapons specifically.

ICRAC suggests one possible difference: the "many hands" problem.[346] A military commander can only be held responsible "for crimes . . . committed by forces under his or her effective command and control."[347] When a war crime is committed by a human soldier, it is normally easy to identify the soldier and determine which military commander (or commanders) had effective control over him. By contrast, when a war crime is "committed" by an autonomous weapon, the number of people involved in the machine's programming and operation can complicate identifying the specific person responsible for the crime. If so—if the specific responsible subordinate is unknown—it may be difficult to identify which military commander (or commanders) had the duty to prevent and punish that crime. And that, in turn, might make it impossible to prosecute any military commander for what the AWS has done.[348]

This is indeed an accountability gap specific to autonomous weapons, but gaps of this type are likely to be exceedingly rare—and possibly non-existent. Although a military commander must have effective control over the individuals who commit a war crime for command responsibility to apply, the International Criminal Tribunal for the Former Yugoslavia (ICTY) explicitly held in *Prosecutor v. Naser Orić* that proving the existence of a superior-subordinate relationship "does not require the identification of the principal perpetrators, particularly not by name . . . provided that it is at least established that the individuals who are responsible for the commission of the crimes were within a unit or a group under the control of the superior."[349] To prosecute a military commander for failing to prevent or punish a war crime committed with an autonomous weapon, therefore, it would not be necessary to identify the specific individual—programmer or operator—responsible for the crime. It would be enough to show that the perpetrator, whoever he is, must have been part of the group over whom the military commander exercised effective control.

When the individual responsible for a war crime committed by an AWS has to be a member of the military, this aspect of command responsibility makes it unlikely that no military commander could be held responsible for the crime. Moreover, because civilian superiors can also be held responsible for crimes

[346] AMOROSO & TAMBURRINI, MEANINGFUL, *supra* note 327, at 7.

[347] Rome Statute, *supra* note 344, art. 28(a).

[348] AMOROSO & TAMBURRINI, MEANINGFUL, *supra* note 327, at 7 (noting that such an outcome "is hardly reconcilable with the moral duty of military commanders").

[349] Prosecutor v. Naser Orić, Case No. IT-03-68-A, App. Ch., Judgment, ¶ 311 (Int'l Crim. Trib. for the Former Yugoslavia June 30, 2006).

committed by their subordinates,[350] an accountability gap is also unlikely to occur when the responsible but unidentifiable individual is civilian—say, a programmer who works for a defense contractor.[351] As long as it is possible to determine that the responsible individual must be a programmer, the corporate official(s) who have effective control over the programmers could be prosecuted via superior responsibility.

To be sure, a *mens rea* requirement applies to military commanders and civilian superiors: the former must have known or "should have known" that his subordinates "were committing or about to commit such crimes" (negligence); the latter must have known or "consciously disregarded information which clearly indicated" that his subordinates "were committing or about to commit such crimes" (recklessness).[352] Similar to direct responsibility, it may be impossible to establish the commander or superior's *mens rea* when one of his subordinates commits a war crime involving an autonomous weapon for the first time. The same is true, however, the first time that a human subordinate commits a war crime. More importantly, once on notice that a war crime *has* been committed by one of his subordinates, it will be much more difficult for a commander or superior to plead lack of *mens rea* if it happens again.[353]

Command responsibility, in short, is likely to lead to an AWS-specific accountability gap only in one very narrow situation: where a human is responsible for a war crime involving an autonomous weapon, that human cannot be identified (precluding direct responsibility), and the group to which the unidentified human belongs is outside of any chain of command, military or civilian (precluding command or superior responsibility). The only relevant difference between that situation and similar situations involving human soldiers is that it might be slightly more difficult to affix direct criminal responsibility to a specific commander when a war crime involves an autonomous weapon. That difference, however, hardly justifies banning AWS.

V. The Jus Ad Bellum

The final consequentialist argument against autonomous weapons focuses on the *jus ad bellum*, not the *jus in bello*: namely, that AWS should be banned because the ability to reduce the number of military casualties "by placing robots

[350] *See, e.g.*, Rome Statute, *supra* note 344, art. 28(b).

[351] *See, e.g.*, Buchan & Tsagourias, *supra* note 321, at 659 ("Programmers outside the chain of command can be treated as commanders themselves under certain circumstances. As we said, the law of command responsibility recognizes de facto authority and control. Moreover, effective command (or authority) and control does not mean exclusive command (or authority) and control.").

[352] Rome Statute, *supra* note 344, art. 28.

[353] *See, e.g.*, Human Rights Watch, *supra* note 21, at 20 ("Actual knowledge of past offenses by a particular set of subordinates may constitute sufficiently alarming information to necessitate further inquiry, and thus may constitute constructive knowledge (reason to know) of future criminal acts satisfying the *mens rea* of command responsibility.").

in harm's way instead of human beings"[354] will make it easier for states to go to war. Sharkey's formulation of this argument is typical:

> Military commanders have a moral responsibility to protect their soldiers. But we must also consider the far-reaching consequences of risk-free war. Having robots to reduce the "body-bag count" could mean fewer disincentives to start wars. In the U.S., since the Vietnam War, body-bag politics has been a major inhibitor of military action. Without bodies coming home, citizens care a lot less about action abroad except in terms of the expense to taxpayers.[355]

As Leonard Kahn notes, the *ad bellum* effect of replacing human soldiers with autonomous weapons is likely to be particularly pronounced in democratic and near-democratic states, because "[t]he loss of a soldier in a democracy or near-democracy is often quite costly in terms of public opinion, and an especially loss-averse electorate can force a state to cease military action after a fairly low number of casualties."[356] This general effect has been amply documented,[357] and a survey in the United States found that "the most common reason given in favour of AWS is 'force protection', i.e. the idea that such weapons can protect the lives of human soldiers."[358]

The *ad bellum* argument is almost certainly descriptively correct, but it is ethically questionable. The argument suggests that international law should ban autonomous weapons because it will force states to risk the lives of as many of their soldiers as possible, thereby maximizing their incentive to avoid going to war. That idea may seem attractive in the context of a war of aggression, where we feel little sympathy for the aggressor. But if deterrence fails, it will not be the aggressor's political and military leaders who will suffer from the absence of autonomous weapons. It will be the human soldiers they send into battle in their place. That

[354] Wagner, *supra* note 112, at 1410–11.

[355] Sharkey, *supra* note 130, at 16; *see also* Sparrow, *supra* note 16, at 106 ("The fact that robotic weapons hold out the prospect of the use of force without risk to one's troops and the likelihood that such systems will be used in more aggressive postures during peace time—again, due to the lack of threat to the life of the "pilot"—suggests that these systems will lower the threshold of conflict and make war more likely."); *Problems with Autonomous Weapons*, *supra* note 11 ("But while replacing people with machines may make military action more politically acceptable at 'home', it can make conflict easier to enter into. It also shifts the burden of harm still further onto civilian populations.").

[356] Leonard Kahn, *Military Robots and the Likelihood of Armed Conflict*, *in* ROBOT ETHICS 2.0: FROM AUTONOMOUS CARS TO ARTIFICIAL INTELLIGENCE 274, 278 (Patrick Lin & Keith Abney eds., 2017).

[357] Robert Johns & Graeme A. M. Davies, *Civilian Casualties and Public Support for Military Action: Experimental Evidence*, 63 J. CONFLICT RESOL. 251, 253 (2019) (noting that "a major literature has built up on the relationship between military casualties and public support for military action").

[358] Nik Hynek & Anzhelika Solovyeva, *Operations of Power in Autonomous Weapon Systems: Ethical Conditions and Socio-Political Prospects*, 36 AI & SOC. 79, 87 (2021).

hardly seems reconcilable with Strawser's "principle of unnecessary risk."[359] Moreover, an AWS ban would affect *any* state using force—not only aggressors, but also states acting in self-defense. Strawser's principle would seem to militate even more strongly against requiring a state like Ukraine, one fighting for its very survival, to put its soldiers in harm's way if "killer robots" could be used instead.

The *ad bellum* argument is also underinclusive. If the threat of soldiers dying in large numbers can undermine a state's willingness to go to war, it is not simply autonomous weapons that are problematic. On the contrary, as even supporters of the argument admit,[360] deterrence concerns would support banning any kind of weapon that limits soldiers' exposure to harm[361]—high-altitude bombing, UAVs, HIMARS, even sniper rifles.

Finally, and most importantly, the *ad bellum* argument is revealingly incomplete. Although public support for the use of force is undermined by the prospect of their soldiers dying, it is also undermined by the prospect of civilian casualties on the other side—a lesson the United States learned the hard way in Vietnam. Simply put, as the most sophisticated research conducted to date indicates, there are only so many My Lais a democratic populace is willing to tolerate:

> [T]he public's casualty aversion extends to foreign civilians. Across two Western public and across multiple experiments with different scenarios, we invariably found support for military action to be lower where the civilian death toll—projected or actual—was higher. The result is robust in another key respect, too, which is that it is strikingly impervious to moderators. We did not find that casualty aversion much weakened when the civilians were from a religious out-group, or when they were described in a less humanizing manner, or when force was more likely to be successful. Rather, support for war falls as civilian casualties increase, largely regardless of whether other things remain the same.[362]

As we have seen, because of their targeting precision and imperviousness to the cognitive, physical, and emotional limits that distort human decision-making, it is likely that autonomous weapons will eventually be able to comply with IHL

[359] Strawser, *supra* note 59, at 344; *see also* Anderson & Waxman, *supra* note 6, at 18 (noting, with regard to the *jus ad bellum* argument, that "to the extent it entails deliberately foregoing available protections for civilians or soldiers in war, for fear that political leaders would resort to war more than they ought, morally amounts to holding those endangered humans as hostages, mere means to pressure political leaders").

[360] *See, e.g.*, Sparrow, *supra* note 16, at 106 (acknowledging that "these sorts of concerns are not specific to AWS and have force against a wider range of means of long-distance war fighting").

[361] *See, e.g.*, Anderson & Waxman, *supra* note 6, at 18 (noting that the argument "can be made with respect to any technological development that either reduces risk to one's own forces or reduces risk to civilians, or both").

[362] Johns & Davies, *supra* note 357, at 270–71.

better than human soldiers, reducing the amount of unnecessary civilian casualties in conflict. A complete version of the *ad bellum* argument, therefore, would not focus solely on how AWS will make it easier for democratic states to go to war by reducing military casualties. It would also emphasize how autonomous weapons will have the same catalyzing effect by producing fewer civilian casualties on the other side than wars fought solely by human soldiers.

That said, even this refined version of the *ad bellum* argument is underinclusive. At this point in their development, autonomous weapons merely promise to reduce the number of foreign civilian casualties. Any currently existing non-autonomous weapon that is more precise than its predecessor has already had that effect—and has thus already made it easier for states to go to war. If AWS critics want to limit war by making it more costly to wage, therefore, they not only have to prohibit further development of killer robots, they must also convince states to renounce most of the weapons they have spent the past century developing. That is a Sisyphean task, to put it mildly, given the powerful incentives states have to develop and use weapons —autonomous and non-autonomous alike—that protect both their own soldiers and civilians on the other side.[363]

CONCLUSION: THE TRAGEDY OF AUTONOMOUS WEAPONS

In the *Hostage* case, decided in the aftermath of World War II, a Nuremberg Military Tribunal held that "[m]ilitary necessity permits a belligerent, subject to the laws of war, to apply any amount and kind of force to compel the complete submission of the enemy with the least possible expenditure of time, life, and money."[364] That quote explains why states are racing to develop autonomous weapons. Simply put, so-called "killer robots" promise to be the ideal instantiation of the principle of military necessity, able to deliver a tremendous amount of force quickly, cheaply, safely, and accurately while still complying with international humanitarian law.

Even the most ardent critics of autonomous weapons accept that they offer significant advantages over human soldiers in terms of firepower, speed, survivability, and accuracy. Their objections to AWS lie elsewhere. For deontological critics, the problem is the sheer inhumanity of autonomous weapons; they would reject humans being killed by machines even if machines complied perfectly with IHL. For consequentialist critics, by contrast, the problem is precisely that they believe autonomous weapons will never be able to comply with IHL as well as human soldiers, making their use ethically and legally impermissible.

[363] *Cf.* Muller, *supra* note 338, at 79 (noting, with regard to autonomous weapons, that "[t]hey make wars less bad, generally, and thus wars are more likely to be chosen as a means").

[364] *United States v. List*, 8 LAW REP. TRIALS WAR CRIMINALS 34, 56 (U.S. Military Tribunal, Nuremberg 1948).

The central argument of deontological critics is not capable of being disproven, because it is an article of faith, not an empirical claim. One either accepts that it is unethical for humans to be killed by a machine or one does not. The central argument of consequentialist critics, however, is explicitly empirical: that autonomous weapons will never be able to comply with IHL as well as human soldiers. One problem with that argument, often mentioned by techno-optimist scholars,[365] is that it is impossible to predict with any certainty how autonomous weapons will develop over time: sensor and AI limits that currently make IHL compliance challenging for machines may not be a problem two decades from now.

But there is an even deeper problem with consequentialist objections to autonomous weapons: namely, that although scholars painstakingly catalog the technological issues that currently limit the ability of AWS to comply with IHL, they show almost no interest in the cognitive issues that have always limited the ability of human soldiers to do so. A fair comparison of machines and humans would consider the limits of both; it would not simply assume that humans are always, or even usually, capable of perceiving the world accurately, understanding rationally, quarantining negative emotions, and reliably translating thought into action. Indeed, as this article has shown, literally decades of research in cognitive psychology demonstrates that human decision-making is profoundly irrational—and never more so than in the heat of combat.

In terms of rational decision-making, then, the non-human is clearly superior to the human. Machines do not rely on the Type 1 heuristic thinking that leads to cognitive biases. Machines do not make physical or mental mistakes because of fatigue or cognitive overload. Machines are impervious to noise, heat, and time-pressure, and they are unaffected by debilitating emotions like stress and anger. And machines do not engage in groupthink. Those advantages do not mean that autonomous weapons are currently capable of complying with IHL as well as or better than human soldiers, nor do they guarantee that AWS will ever have that capacity. Nevertheless, given that autonomous technology is far more likely to improve than human decision-making, it is difficult to exclude the possibility that, in time, autonomous weapons will indeed be able to outperform human soldiers in terms of IHL compliance.

In light of this possibility, why are consequentialist critics so quick to insist that autonomous weapons be banned—especially those who, like Sparrow and Heyns, believe that the use of AWS would be ethically acceptable, perhaps even ethically required, if they could comply with IHL as well as human soldiers?

As this article has shown, there are two very different answers to that question. The first is that some consequentialists simply believe "killer robots" will

[365] Lena Trabucco & Kevin Jon Heller, *Beyond the Ban: Comparing the Ability of Autonomous Weapon Systems and Human Soldiers to Comply with IHL*, 46 FLETCHER F. WORLD AFF. 15, 10 (2022).

always be worse soldiers than humans, whether because they are pessimistic about AWS technology, optimistic about the judgment of human soldiers, or both. The second answer, and the more interesting one, is that some consequentialists fear that killer robots will actually be *better* soldiers than humans, not worse. These consequentialists do not want more lawful violence and less unlawful violence, which is the promise (however distant) of IHL-compliant autonomous weapons. They want less violence of any kind— unlawful *and* lawful. And that is only possible insofar as wars continue to be fought by human soldiers.

The fear that AWS will be better soldiers than humans is evident in arguments that focus on compassion. As discussed earlier, critics believe that the ability of soldiers to show compassion is ethically required not only to protect civilians against unnecessary harm, but also to leave open the possibility that a soldier will not kill the enemy even when he is lawfully entitled to do so. Most soldiers do not hesitate to follow their training, but some will still stay their hand, even in the face of mortal danger. Killer robots, by contrast, have no such compunctions. Because machines do not "decide" whether to pull the trigger, no autonomous weapon will hesitate to fire when it encounters an individual who matches the targeting parameters established by its programming.[366]

For the reasons discussed in Section III, the compassion argument is unpersuasive. The fear that autonomous weapons will be better soldiers than humans is also evident, however, in the ethically problematic but descriptively convincing *jus ad bellum* argument. If the decision to go to war in democratic states is influenced by the expected number of military losses on their side and civilian losses on the other side, which is what research indicates, the superiority of AWS to human soldiers in both respects may well make war more common. It is highly likely that AWS will eventually comply with IHL better than human soldiers—and people on the home front do not care when robots come home in pieces. Humane war fought by machines could thus all too easily lead to the kind of "endless war" that Samuel Moyn has so eloquently decried.[367]

Sir Peter Shaffer, the great English playwright, once described tragedy as "not a conflict between right and wrong, but between two different kinds of right."[368] That quote encapsulates the dilemma created by autonomous weapons—and by any weapon that protects both combatants and civilians from needless harm. It is right to make war more humane, because unnecessary death and suffering in conflict should always be avoided. But it is even more right to end war itself, because without war there would be no death or suffering to minimize. Those two

[366] *See, e.g.*, BOULANIN ET AL., *supra* note 269, at 13 ("An AWS applies force when the data received as input from its sensors matches the parameters of the target profile.").

[367] *See generally* SAMUEL MOYN, HUMANE: HOW THE UNITED STATES ABANDONED PEACE AND REINVENTED WAR (2022).

[368] *Peter Shaffer Quotes*, QUOTE.ORG, https://quote.org/quote/tragedy-for-me-is-not-a-conflict-482908 [https://perma.cc/V8TM-8LC6].

ideals, however, cannot be reconciled: the more humane war becomes, the more difficult it will be to eradicate it. That is the choice, and only humans can make it.