

JUSTICE, HEALTH, AND DEMOCRACY
IMPACT INITIATIVE & CARR CENTER
FOR HUMAN RIGHTS POLICY

Handle With Care

Autonomous Weapons and Why the
Laws of War Are Not Enough



Technology & Democracy
Discussion Paper

Handle With Care

Autonomous Weapons and Why the Laws of War Are Not Enough

Linda Eggert

Technology and Human Rights Fellow
Carr Center for Human Rights Policy
Harvard Kennedy School

Series edited by **Joshua Simmons**

Technology and Human Rights Fellow
Carr Center for Human Rights Policy
Harvard Kennedy School

and

Postdoctoral Fellow in Technology and Democracy
Justice, Health & Democracy Impact Initiative
Edmond & Lily Safra Center for Ethics
Harvard University

September 2022 | Issue 2022-07

For helpful comments and discussion, the author is grateful to Gabriella Blum, Daniel Butt, John Emery, Seth Lazar, Mathias Risse, Thomas Simpson, Victor Tadros, Miles Unterreiner, and an anonymous reviewer, as well as audiences at Harvard's Carr Center for Human Rights Policy, Stanford's Center for Ethics in Society, the University of Zurich, and Hamburg University. Special thanks to Josh Simons and Eli Frankel for their efforts, as well as to Barbara Smith-Mandell for copy-editing.



HARVARD Kennedy School

CARR CENTER
for Human Rights Policy

The Justice, Health, and Democracy Impact Initiative is a multi-institutional collaboration between New America, the Brown University School of Public Health, and the Edmond & Lily Safra Center for Ethics at Harvard University.

Cover Image credit: Sarah Holmlund, Adobe Stock; Image p. 7 credit: diablosiatr, Adobe Stock

A common view is that autonomous weapon systems (AWS) should be banned because they might violate the laws of war. But even perfect compliance with the law is compatible with morally wrongful killing on a massive scale. As a criterion for assessing the ethics of deploying AWS, therefore, the law must be handled with care.

INTRODUCTION

In spring 2013, a global coalition, the Campaign to Stop Killer Robots, launched with a mission to advocate for a ban on “machines that determine whom to kill.”¹ Nine years later, almost to the day at the time of writing, no such ban exists. Autonomous weapons research is alive and well, and artificial intelligence has made it to the fore of the Pentagon’s future weapons development strategy. The latest Review Conference of the Convention on Conventional Weapons (CCW), a primary forum for international talks on lethal autonomous weapon systems, failed to achieve consensus on whether new international laws are needed to address threats posed by autonomous weapons technology.² Meanwhile, high-tech military powers, including China, Russia, Israel, South Korea, the US, and the UK, continue to invest heavily in the development of autonomous weapon systems.

An autonomous weapon system (AWS), according to the International Committee of the Red Cross’s working definition, is

[a]ny weapon system with autonomy in its critical functions. That is, a weapon system that can select (i.e., search for or detect, identify, track, select) and attack (i.e., use force against, neutralize, damage or destroy) targets without human intervention.³

In other words, AWS automate the critical functions of selecting and engaging targets. To this extent, autonomy in AWS correlates with a lack of human involvement in decisions about targeting and the use of

force. This article’s primary concern is with AWS that target humans.

Calls for a ban on “killer robots” have received support from roboticists, scholars, activists, Nobel peace laureates, and others.⁴ Reasons include an array of security risks as well as intrinsic objections, given robots’ lack of empathy, mercy, moral judgment, and understanding of the value of human life. One especially widely shared worry is that AWS may not be able to comply with the laws of armed conflict (also known as international humanitarian law, or IHL). In roboticist Ron Arkin’s words, AWS “must be constrained to adhere to the same laws as humans or they should not be permitted on the battlefield.”⁵

This paper warns that, though seemingly natural and ubiquitous, appeals to IHL should be handled with care. Even if AWS could be made never to violate IHL, this would be no sound indication that their use is morally permissible. For while IHL is essential to reducing violence in armed conflict, abiding by IHL may nonetheless be compatible with morally wrongful killing on a massive scale.

This warning differs from both contingent objections to AWS, which typically focus on technology’s limits, and intrinsic objections, such as that “killing by algorithm” is such an affront to human dignity that it could never be morally permissible. By interrogating compliance with IHL as a criterion for assessing the moral permissibility of deployment, this paper illuminates an altogether different dimension of the debate: what criteria

1 Stop Killer Robots, “Our Vision and Values,” <https://www.stopkillerrobots.org/vision-and-values/> (accessed 21 July 2022).

2 Emma Farge, “U.N. Talks Adjourn Without Deal to Regulate ‘Killer Robots,’” *Reuters*, 17 Dec. 2021, <https://www.reuters.com/article/us-un-disarmament-idAFKBN2IW1UJ>.

3 ICRC, “Views of the International Committee of the Red Cross (ICRC) on Autonomous Weapons Systems,” 11 April 2016, Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), 11–15 April 2016, Geneva, <https://www.icrc.org/en/download/file/21606/ccw-autonomous-weapons-icrc-april-2016.pdf>. For additional background, see United Nations Office of Disarmament Affairs, “Background on Laws in the CCW,” <https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/>.

4 Future of Life Institute, “Autonomous Weapons: An Open Letter from AI & Robotics Researchers,” 28 July 2015, <https://futureoflife.org/2016/02/09/open-letter-autonomous-weapons-ai-robotics/>; Nobel Women’s Initiative, “Nobel Peace Laureates for Preemptive Ban on ‘Killer Robots,’” 12 May 2014, <https://www.nobelwomensinitiative.org/nobel-peace-laureates-call-for-preemptive-ban-on-killer-robots/>; Ariel Conn, “An Open Letter to the United Nations Convention on Certain Conventional Weapons,” Future of Life Institute, 20 Aug. 2017, <https://futureoflife.org/2017/08/20/autonomous-weapons-open-letter-2017/>.

5 Ronald C. Arkin, *Governing Lethal Behavior in Autonomous Robots* (New York: CRC Press, 2009), 33.

we should apply to begin with, as we confront the moral and legal conundrums of the increasing *autonomization* of warfare.⁶

Whether AWS could abide by the current laws of armed conflict is an inadequate test for whether it is morally permissible for militaries to deploy them. Harms lawfully caused might be morally wrong. While there are good reasons to legally permit combatants to harm morally innocent people (that is, people who have rights not to be harmed), these reasons do not extend to robots. Instead of relying on IHL as a basis for assessing the ethics of autonomized conduct in war, we should consider the moral demands imposed by individual human rights. These, it turns out, raise altogether different questions.

The purpose of this short piece is not to take an all-things-considered stance on whether the international community should impose a ban on lethal AWS but, above all, to caution against extending human combatants' legal permissions to acts performed by AWS. This throws new light on how AI's expanding role in war might require us to confront the moral limits inherent in our current laws.

What follows is, first, a sketch of a contemporary worry that applies to IHL in general; second, an argument that this worry is especially serious if applied to AWS; and, third, a starting point for rethinking the challenge of regulating the autonomized conduct of armed conflict.

MORALITY AND THE LAW

IHL—international humanitarian law—governs the conduct of hostilities in armed conflict. Comprising the Geneva Conventions and their Additional Protocols, as well as other conventions and customary international law, IHL protects persons not directly participating in hostilities, that is, non-combatants and those *hors de combat*.⁷

A key principle of IHL is that of distinction. It requires that the parties to an armed conflict at all times distinguish between combatants and civilians, and military and civilian objects.⁸ Intentional attacks against civilians and civilian objects are prohibited. A common fear is that AWS might not be able to distinguish between

combatants and non-combatants, and so would violate the principle of distinction.⁹ For example, to a robot, a child waving a toy gun might be indistinguishable from a combatant with a real weapon.¹⁰

A related worry concerns the principle of proportionality: this essentially prohibits attacks in which the foreseen harm to civilians is excessive in relation to the anticipated military advantage.¹¹ How could the task of assessing whether foreseeable harm to civilians morally outweighs the anticipated military advantage possibly be algorithmically represented?

Principles like distinction and proportionality embody significant advances in the international community's efforts to limit the harmful effects of armed conflict—not least by prohibiting intentional attacks against civilians. But complying with IHL is compatible with killing morally innocent people. This should give us pause as we deliberate whether compliance with IHL provides a sufficient standard in assessing the moral permissibility of deploying AWS. Before we proceed, note that the relationship between the ethics and law of war is the subject of long-running scholarly debates. We are merely dipping our toes in the water here.

Article 51(5) of Additional Protocol I of the 1977 Geneva Conventions prohibits any “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.”¹²

Besides the fact that there is significant room for interpretation of what harms count as “excessive,” note that the advancement of military objectives takes precedence over persons' rights not to be killed. The law subordinates individual rights not to be killed to the advancement of even unjust war aims. By restricting the unintended killing of civilians only to the extent that this would outweigh the anticipated military advantage, IHL effectively determines the permissibility of harms to civilians purely in relation to the advancement of military goals, irrespective of the justness of the cause. Currently, for example, this means that Russian combatants are acting within their legal rights in killing civilians in Ukraine, so long as these casualties

6 The question is neither whether deploying AWS is compatible with IHL nor whether harms inflicted by AWS are in keeping with IHL. The question is to what extent robots' compliance with IHL serves as a reliable indication of the moral permissibility of their use.

7 For discussion of IHL's main principles, see M. Cherif Bassiouni, “The Normative Framework of International Humanitarian Law: Overlaps, Gaps, and Ambiguities,” *Transnational Law & Contemporary Problems* 8 (1999): 199–200; Jean-Marie Henckaerts and Louise Doswald-Beck, *Customary International Humanitarian Law*, vol. 1, *Rules* (Cambridge: Cambridge University Press/ICRC, 2005, reprinted with corrections 2009); Adam Roberts and Richard Guelff, eds., *Documents on the Laws of War*, 3rd ed. (Oxford: Oxford University Press, 2000).

8 Combatants “shall at all times distinguish between [civilians or] civilian objects and military objectives and accordingly shall direct their operations only against military objectives” (Protocol I, Article 48). See also Henckaerts and Doswald-Beck, *Customary International Humanitarian Law*, vol. 1, *Rules*, 3.

9 Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), adopted 8 June 1977, 1125 U.N.T.S. 3, entered into force 7 December 1978, art. 51(3); Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of Non-International Armed Conflicts (Protocol II), 1125 U.N.T.S. 609, entered into force 7 December 1978, art. 13(3).

10 For discussion, see Alex Leveringhaus, *Ethics and Autonomous Weapons* (Oxford: Palgrave Macmillan, 2015), 345..

11 Protocol I, art. 51(5).

12 “1977 Geneva Protocol I Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts,” in Adam Roberts and Richard Guelff, eds., *Documents on the Laws of War* (Oxford: Oxford University Press, 1982), 416.



Credit: Stop Killer Robots/Ralf Schlesener

are not intended, and are deemed “necessary” and “proportionate” to the anticipated military advantage Russia thereby gains.

From a moral perspective, this is absurd. No harm is “proportionate” to achieving a goal that is itself unjust. Nonetheless, for reasons to which we will shortly come, it is almost universally agreed that the law must grant all combatants, no matter on what side they are, the same permission to use lethal force, including if this will harm civilians as a side effect.

The same permission to use lethal force applies to all parties to an armed conflict. Once a war is underway, there is no legal difference between attacks carried out in pursuit of a just cause and attacks carried out in pursuit of an unjust cause. So long as combatants distinguish themselves from civilians, they may “participate directly in hostilities,” irrespective of the moral nature of each party’s war aim.¹³ One upshot of this is that the law permits the targeting of combatants, even if they have done nothing to lose their moral right not to be harmed. This is why Russian invaders enjoy the same legal permission to use lethal force as Ukrainians defending their country against the unjust invasion.

In these respects, IHL effectively permits acts of killing that morality prohibits: it does not prohibit “collateral” harms that are deemed necessary and proportionate, even in pursuit of an unjust cause; and it does not prohibit the targeting of combatants who have not done anything to lose their moral rights not to be harmed. Thus, in many cases, the law does not distinguish between *morally* permissible and impermissible uses of lethal force.

There are good reasons for this. For example, it is often difficult to determine with absolute certainty whether a cause is just, so there are prudential reasons against legally distinguishing between a “just” and “unjust” side. It is also often practically impossible for

combatants to judge with accuracy which individuals might have moral rights not to be harmed. Battlefields are hardly conducive to careful thought. Moreover, some combatants fighting for an unjust aim may be acting under duress or have been misled into believing that they are fighting for a just cause. Consider, for example, the Russian combatants who were duped into believing they would be liberating people in Ukraine from “fascist monsters.”¹⁴

Not all combatants are blameworthy for doing what is morally wrong, and it would be unfair to prohibit non-blameworthy people from defending their lives. This is one of the reasons why it makes sense for the law to permit all combatants to use lethal force—even at the cost of failing to prohibit morally wrongful killing. It is, moreover, a longstanding convention that combatants are not held responsible for the justness of their war; and any challenge to this, one might fear, would destabilize an already all-too-fragile system. These considerations are of no concern here.

THE LAW’S HUMAN-CENTEREDNESS

The moral logic, according to which humans may be absolved from blame for doing what is morally wrong, and which substantiates a law that does not prohibit the killing of morally innocent people, applies only to human moral agents. It does not apply to robots. Robots are not the kinds of entities that could be innocent or culpable, nor do they possess characteristics that might justify not legally prohibiting acts of killing that are morally impermissible.

The law is appropriately sensitive to how human combatants experience the battlefield, including the fact that duress and epistemic constraints make it extremely difficult, if not impossible, to abide by the demands of morality. But there is no comparable sense in which AWS “experience” the battlefield.

¹³ Protocol I, 1125 UNTS 3, Articles 43 (2), 44; International Committee of the Red Cross, *Customary International Humanitarian Law*, vol. 1, Rules.

¹⁴ Luke Harding, “Demoralised Russian Soldiers Tell of Anger at Being ‘Duped’ into War,” *The Guardian*, 4 March 2022, <https://www.theguardian.com/world/2022/mar/04/russian-soldiers-ukraine-anger-duped-into-war>

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. While it makes sense for the law to make special provisions for humans, many of whom are at least partially excused for committing moral wrongs, it does not make sense to extend the legal permission to commit morally wrongful acts of killing to AWS.

autonomized harming is permitted. For the purposes of this discussion, what matters is that, to assess and regulate autonomized harming in war in the first place, we would need a set of principles that are sensitive to individual rights. This is not the case for IHL.

HUMAN CONTROL AND HUMAN RIGHTS

Where does this leave leaders around the world who face calls for a new international framework to “ensure

"Perhaps the growing individualization and autonomization of war calls for a greater degree of individualization in the law."

But, one might ask, if there are sound reasons for legally permitting combatants to kill morally innocent people, then why should this change with the types of weapons militaries use? If the law is right not to prohibit lethal harms to morally innocent people, why shouldn't this extend to the use of AWS? Shouldn't the same rationale for legally permitting some morally wrongful acts of killing also apply to the deployment of AWS? After all, AWS are mere weapons, and might be deployed in the mistaken belief that the cause is just, without any intention of targeting civilians, and after assessing that any harm that might be caused to civilians would not be disproportionate to the anticipated military advantage.

But even if IHL serves as a morally adequate legal framework for governing human combatants' conduct in war, this does not mean that IHL also serves as a morally adequate legal framework for governing autonomized conduct. Robots tasked with identifying and engaging targets should not be governed by a law that permits the killing of innocent people in the absence of a sound moral justification. The fact that we can morally justify a law that does not prohibit human combatants from inflicting lethal harms on morally innocent people does not automatically entail that we can morally justify the autonomized harming of morally innocent people.

A key reason for banning AWS, then, is not just that they might violate the principles of IHL, but that even harms lawfully caused might be morally impermissible. While we have good reasons not to legally prohibit human combatants from doing what is morally wrong, these reasons do not apply to robots.

A more comprehensive discussion would ultimately need to establish, first, what constitutes just autonomized harming and, second, which principles accordingly govern the just employment of AWS by human agents. The former question takes precedence over the latter: just employment of AWS is only a possibility if just

that technology is developed and used to promote peace, justice, human rights, equality and respect for law"¹⁵

Several states, including the US, Russia, and India, continue to resist calls for negotiations of a new legally binding instrument to regulate the development and use of AWS, insisting that existing IHL is sufficient.¹⁶ This position is flawed. Algorithmic capacity to follow existing IHL is an inadequate standard for assessing the moral permissibility of deploying AWS.

Our task goes far beyond clarifying how IHL, as we know it, might apply to AWS. Nor is it merely to maintain a certain level of “meaningful human control” over the use of force, to ensure compliance with IHL's requirements of distinction, proportionality, and precaution. (See, for example, [Venezuela's Closing Statement](#) on behalf of the Non-Aligned Movement at the CCW last autumn.) Human control might help address fears about a potential “responsibility gap” for harms inflicted through AWS. But even with meaningful human control, the issue highlighted here would persist. Ensuring, through human control, that AWS would not violate existing IHL might still, without a sound justification, legally allow acts of killing that are morally impermissible.

If capacity to follow IHL is not an adequate criterion for assessing the moral permissibility of harms caused by AWS, the legal principles that should govern AI systems' conduct in armed conflict will need to differ significantly from the legal principles that currently govern human combatants' conduct. Rather than asking merely whether AWS would be able to follow traditional legal rules enshrined in IHL, we should ask whether AWS would be able to spare people who have moral rights not to be harmed.

This is not to say that we should eliminate the combatant/civilian distinction in the law. The legal prohibition on attacking civilians

15 Isabelle Jones, “GGE Pushes Decisions to Critical Review Conference,” Stop Killer Robots: News, 12 Sept. 2021, <https://www.stopkillerrobots.org/news/gge-pushes-decisions-to-critical-review-conference/>.

16 “[W]e believe the existing IHL and effective measures at the national level to implement IHL are sufficient to address the challenges posed by LAWS,” quoted in HRW and IHRC, “Crunch Time on Killer Robots: Why New Law Is Needed and How it Can Be Achieved,” 5n12.

is not a principle we should jeopardize. Rather, the point is that our concern should be not just with human control exercised by human *agents*, but also with the human rights of human *patients*—those who might suffer harms. Preserving human *agency* through “meaningful human control” is not enough. We need to keep sight of what else is at stake: the rights of those who might be harmed.

One natural starting point is the international human rights framework. Although the relationship between IHL and international human rights law (IHRL) is fiercely contested, it is widely acknowledged that IHRL at least better tracks morally individualist commitments than IHL.

An additional reason for focusing on human rights when considering the legal regulation of AWS is that IHRL applies in a wider range of contexts than IHL, and the use of AWS may not be limited to armed conflicts. Other situations in which AWS might be used include law enforcement and counterterrorism efforts. As Sri Lanka

put it, applying IHRL to AWS is “logical and pertinent” precisely because AWS could be used in situations outside of armed conflict.¹⁷

It would go far beyond the scope of this piece to discuss the relationship between IHL and IHRL. The suggestion for further discussion is that any new legally binding instrument governing AWS should go far beyond existing IHL to protect the rights of individuals more effectively than IHL is designed to do.

WHAT’S THE POINT?

What about the possibility that AWS are such an affront to human dignity that their use is inherently wrong? If “killer robots” are *mala in se*, bad in themselves, is there any point in asking what rules they should follow?

As of now, it looks like the use of AI in war will only increase. If the conduct of hostilities continues to become more autonomized, we had better have thought carefully about what rules should be in place in these highly non-ideal circumstances – where people may not do what they ought to do. Just like it is sensible to have laws that limit the destructive force of wars, even though unjust wars should not be fought at all, it may be sensible to devise laws that prohibit at least certain acts by AWS, even if AWS should not be used at all.

This article offers opponents of AWS a basis on which to mount their objections that is more robust than the worry that AWS might not be able to abide by the rules of IHL. If IHL is insufficient to begin with, treating it as a main moral criterion in assessing the ethics of using AWS would be ill-considered. And if it turns out that AWS cannot possibly be made to comply with more restrictive, rights-based rules that *would* apply to them, then the case against deployment is all the more urgent.

WHERE DO WE GO FROM HERE?

In the absence of an outright ban on AWS, one possibility is that we end up with two bodies of law: IHL for human combatants and a more individualized, restrictive set of principles for AWS. In this scenario, IHL continues to account for the horrendous circumstances of war that make it practically impossible for people to avoid committing moral crimes, and is accordingly permissive when it comes to the targeting of potentially morally innocent people; while the law applying to AWS permits harms only under circumstances in which human targets are overwhelmingly likely morally liable to be harmed. From a policy perspective, devising different laws for different types of weapons seems impossibly impracticable, to say nothing of the notorious difficulties



¹⁷ Statement of Sri Lanka, CCW GGE Meeting on Lethal Autonomous Weapons Systems, Geneva, 29 September 2021 (UN audio files), http://149.202.215.129:8080/s2t/UNOG/LAWS-29-09-2021-AM_mp3_en.html. For what it’s worth, the ICRC and several states at the GGE have also recognized the relevance of IHRL to governing the use of AWS. Argentina, Austria, Brazil, Chile, Costa Rica, Mexico, New Zealand, Palestine, Panama, and the Philippines appealed to the relevance of IHRL; while Israel and India opposed such reference to IHRL; see HRW and IHRC, “Crunch Time on Killer Robots,” 13n48.

of determining individual moral liability, defining autonomy in weapon systems, and distinguishing different degrees of it.

Another—arguably more desirable—possibility would be to rethink the relationship between IHL and the international human rights framework in general, and to ascribe a more decisive role to human rights protection in armed conflict than it has traditionally been afforded. IHL's rootedness in traditional, collectivist notions appears increasingly at odds with the changing character of war: not only may the conduct of armed conflict become increasingly autonomized, but precision-guided technologies already increasingly facilitate the targeting of individuals. Perhaps the growing individualization and autonomization of war calls for a greater degree of individualization in the law.

As long as a legal framework geared towards the protection of individual rights remains outside the reach of feasibility, one option is to prohibit the autonomized targeting of human persons altogether, and to limit the use of AWS to the targeting of physical property.¹⁸ Another is to permit only the autonomized infliction of non-lethal harms, such as by incapacitating people through sound waves or non-blinding lasers. Both are examples of applications already in use.¹⁹

TO CLOSE

According to one of the guiding principles agreed by the UN Group of Governmental Experts on Lethal AWS, it is critical to ensure that the potential use of AWS “is in compliance with applicable international law, in particular IHL.”²⁰ More recently, Austria, Brazil, Chile, Ireland, Luxembourg, Mexico, and New Zealand wrote in a joint submission to the GGE that we should “ask not only if a weapon is legally acceptable (can the weapon be used in accordance with the law?) but, would its use be acceptable from an ethical perspective: (should we use this weapon?).”²¹

This article sought to illuminate the gulf between what is “legally acceptable” and what is “acceptable from an ethical perspective”

as we grapple with the moral complexities that accompany the increasing autonomization of armed conflict. This article maintained that, contrary to common assumptions, AI systems' ability to abide by the current laws of war offers no reliable indication as to whether deployment would be morally permissible, because even perfect compliance with IHL may be compatible with morally wrongful killing on a massive scale. Rather than asking whether AWS would be able to abide by traditional legal norms, we should ask whether they would be able to act in accordance with people's rights not to be harmed. These are different questions.

AI's expanding role in war presses us to confront the moral limits inherent in our current laws. The fact that IHL permits morally wrongful acts of killing should be at the fore of policy debates about what rules should govern AI systems in armed conflict, and specifically about whether IHL is sufficient. While we have good reasons to legally permit combatants to use lethal defensive force, even if they might kill people who have rights not to be killed, we should not extend the legal permission to commit morally wrongful acts of killing to AWS.

Since there is currently no existing international law dedicated specifically to AWS, what is known as the Martens Clause has acquired special relevance in debates about AWS.²² The Martens Clause resides within Additional Protocol I to the Geneva Conventions, and applies in situations not specifically covered by international agreements and mandates, such that “civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from the dictates of public conscience.”²³ Whatever else the “principles of humanity” and “dictates of public conscience” might demand, legal exceptions to the moral prohibition on killing the innocent must be handled with the utmost care, not least when human rights are at stake and AI is in charge.

18 William F. Schulz and Sushma Raman, *The Coming Good Society: Why New Realities Demand New Rights* (Cambridge, MA: Harvard University Press, 2020), 203.

19 JNLWP, “Non-Lethal Weapons (NLW) Reference Book,” Joint Non-Lethal Weapons Directorate, Quantico, VA, 2012, https://www.supremecourt.gov/opinions/URLs_Cited/OT2015/14-10078/14-10078-3.pdf.

20 Final Report, Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Guiding Principles affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/MSP/2019/CRP.2/Rev.1, November 2019, Annex III, para. (c).

21 “Submission by Austria, Brazil, Chile, Ireland, Luxembourg, Mexico and New Zealand on Ethical Considerations to the Chair of the Group of Governmental Experts on Technologies in the Area of Lethal Autonomous Weapons Systems” (2021), available at <http://www.converge.org.nz/pma/jsub-gge,sept21.pdf>.

22 For example, Stephen Goose, “The Case for Banning Killer Robots,” *Association for Computing Machinery* 58 (2015): 43–45.

23 Protocol I, art. 1(2).

Technology and Democracy Discussion Paper Series

**Justice, Health, and Democracy Impact Initiative &
Carr Center for Human Rights Policy
Harvard University
Cambridge, MA 02138**

Statements and views expressed in this paper are solely those of the author and do not imply endorsement by Harvard University, the Justice, Health, and Democracy Impact Initiative, or the Carr Center for Human Rights Policy.

Copyright 2022, President and Fellows of Harvard College
Printed in the United States of America

**This publication was published by the
Justice, Health, and Democracy Impact Initiative &
Carr Center for Human Rights Policy
at Harvard University**

Copyright 2022, President and Fellows of Harvard College
Printed in the United States of America



**carrcenter.hks.harvard.edu
ethics.harvard.edu/JHD-impact-initiative**

79 JFK Street | Cambridge, MA 02138
617.495.5819