Volume 94 Number 886 Summer 2012

INTERNATIONAL REVIEW of the Red Cross

On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making

Peter Asaro

Prof. Asaro is a philosopher of technology who has worked in artificial intelligence, neural networks, natural language processing, and robot vision research. He is an Affiliate Scholar at Stanford Law School's Center for Internet and Society, Co-Founder and Vice-Chair of the International Committee for Robot Arms Control, and the Director of Graduate Programs for the School of Media Studies at The New School for Public Engagement in New York City.

Abstract

This article considers the recent literature concerned with establishing an international prohibition on autonomous weapon systems. It seeks to address concerns expressed by some scholars that such a ban might be problematic for various reasons. It argues in favour of a theoretical foundation for such a ban based on human rights and humanitarian principles that are not only moral, but also legal ones. In particular, an implicit requirement for human judgement can be found in international humanitarian law governing armed conflict. Indeed, this requirement is implicit in the principles of distinction, proportionality, and military necessity that are found in international treaties, such as the 1949 Geneva Conventions, and firmly established in international customary law. Similar principles are also implicit in international human rights law, which ensures certain human rights for all people, regardless of national origins or local laws, at all times. I argue that the human rights to life and due process, and the limited conditions under which they can be

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

doi:10.1017/S1816383112000768

overridden, imply a specific duty with respect to a broad range of automated and autonomous technologies. In particular, there is a duty upon individuals and states in peacetime, as well as combatants, military organizations, and states in armed conflict situations, not to delegate to a machine or automated process the authority or capability to initiate the use of lethal force independently of human determinations of its moral and legal legitimacy in each and every case. I argue that it would be beneficial to establish this duty as an international norm, and express this with a treaty, before the emergence of a broad range of automated and autonomous weapons systems begin to appear that are likely to pose grave threats to the basic rights of individuals.

Keywords: robots, drones, autonomous weapon systems, automation, lethal decision-making, human rights, arms control.

: : : : : : :

In September 2009, the International Committee for Robot Arms Control (ICRAC)[1] was formed by Jürgen Altmann, Noel Sharkey, Rob Sparrow, and me. Shortly thereafter we issued a mission statement that included a call for discussion about the establishment of an international prohibition on autonomous weapon systems:

> Given the rapid pace of development of military robotics and the pressing dangers that these pose to peace and international security and to civilians in war, we call upon the international community to urgently commence a discussion about an arms control regime to reduce the threat posed by these systems. We propose that this discussion should consider the following: The prohibition of the development, deployment and use of armed autonomous unmanned systems; machines should not be allowed to make the decision to kill people.[2]

Since then, the issue has been taken up by philosophers, legal scholars, military officers, policymakers, scientists, and roboticists. The initial discussion has focused on the inability of existing autonomous weapon systems to meet the legal requirements of international humanitarian law (IHL), and conjectures as to the possibility that future technologies may, or may not, be able to meet these requirements. Of particular concern has been whether autonomous systems are capable of satisfying the principles of distinction and proportionality required by the Geneva Conventions, and whether it will be possible to hold anyone responsible for any wrongful harms the systems might cause. On the basis of the initial discussions, attention has begun to turn to the question of whether IHL needs to be

1 See www.icrac.net.

2 Jürgen Altmann, Peter Asaro, Noel Sharkey and Robert Sparrow, *Mission Statement of the International Committee for Robot Arms Control*, 2009, available at: http://icrac.net/statements/ (this and all links last visited June 2012).

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

supplemented with an international treaty that explicitly prohibits these technologies. While the vast majority of people and a number of scholars, lawyers, military officers, and engineers agree that lethal systems should not be autonomous, there are some who take the position that an international prohibition on autonomous weapon systems may be premature, unnecessary, or even immoral.[3] I believe that this latter position is mistaken, and propose that we must act soon to prohibit these systems. I will argue that we have moral and legal duties to prevent the delegation of lethal authority to unsupervised non-human systems, and to invest our science and engineering research and development resources in the enhancement of the ethical performance of human decision-makers. To support this argument, this article will supply a theoretical foundation for an international ban on autonomous weapon systems based on international human rights law (IHRL) and IHL. In addition to being enshrined in and protected by a large body of international and domestic law, human rights also have a moral status independent of existing law, and thus can provide sound guidance for the extension of the law to deal with the issues raised by emerging technologies. I will argue that an international ban on autonomous weapon systems can be firmly established on the principle that the authority to decide to initiate the use of lethal force cannot be legitimately delegated to an automated process, but must remain the responsibility of a human with the duty to make a considered and informed decision before taking human lives.

This principle has implications for a broad range of laws, including domestic laws, IHRL, and IHL. Insofar as the current interest in developing autonomous weapon systems is motivated primarily by military applications, I will focus on the IHL implications. However, the same principle would apply to the use of autonomous weapon systems by states for domestic policing, crowd control, border control, guarding prisoners, securing facilities and territory, or other potentially lethal activities, as well as to their use by individuals or organizations for a broad range of security applications involving the use of force. Similarly, I will focus on the human right to life, though similar arguments might be made regarding automated decisions to override or deny other human rights, in automating activities such as: arrest, detention, and restriction of movement; search, surveillance and tracking; deportation; eviction and foreclosure; denial of healthcare, public assembly, freedoms of press and speech, voting rights; and other civil, political, economic, social, and cultural rights.[4]

3 Ronald C. Arkin, *Governing Lethal Behavior in Autonomous Robots*, CRC Press, 2009; Gary Marchant, Braden Allenby, Ronald C. Arkin, Edward T. Barrett, Jason Borenstein, Lyn M. Gaudet, Orde F. Kittrie, Patrick Lin, George R. Lucas, Richard M. O'Meara and Jared Silberman, 'International governance of autonomous military robots', in *Columbia Science and Technology Law Review*, 30 December 2010, available at: http://ssrn.com/abstract=1778424; Kenneth Anderson and Matthew C. Waxman, 'Law and ethics for robot soldiers', in *Policy Review*, 28 April 2012, available at: http://ssrn.com/abstract=2046375.

4 The human rights currently recognized in international law include, but are not limited to, the rights enshrined in the United Nations International Bill of Human Rights, which contains the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights, and the International Covenant on Economic, Social and Cultural Rights.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

Autonomous weapon systems

Recent armed conflicts have seen an increased use of highly automated technologies, the most conspicuous being the use of armed, remotely piloted drones by the US military (among others) in a number of countries. These combat aircraft are capable of numerous sophisticated automated flight processes, including fully automated take-off and landing, GPS waypoint finding, and maintaining an orbit around a GPS location at a designated altitude, as well as numerous automated image collection and processing capabilities. While these systems are highly *automated*, they are not considered to be *autonomous* because they are still operated under human supervision and direct control.[5] Moreover, despite being armed with weapons that have some automated capabilities, such as laser-guided missiles and GPS-guided bombs, these systems still rely on direct human control over all targeting and firing decisions. The crucial concern of this article is with the legal and ethical ramifications of automating these targeting and firing decisions. We can thus define an 'autonomous weapon system' as any system that is capable of targeting and initiating the use of potentially lethal force without direct human supervision and direct human involvement in lethal decision-making.[6] Under this definition, current remote-piloted aircraft, such as Predator and Reaper drones, are not autonomous weapon systems. However, it is becoming increasingly clear that those activities that currently remain under human control might be automated in the near future, making possible the elimination of direct human involvement in target selection and decisions to engage targets with lethal force. Remote-piloted aircraft are not the only concern, as there are now numerous land, sea, and submarine systems that might also be armed, as well as fixed defensive systems, such as gun turrets and sentries, and various modes of cyber attack, which might be similarly automated so as to be capable of delivering lethal force without the direct involvement of human beings in selecting targets or authorizing the use of lethal force against a target.

While there are various examples of military weapons and practices that, arguably, do not include direct human involvement in lethal decision-making, this

5 The term 'autonomous' is used by engineers to designate systems that operate without direct human control or supervision. Engineers also use the term 'automated' to distinguish unsupervised systems or processes that involve repetitive, structured, routine operations without much feedback information (such as a dishwasher), from 'robotic' or 'autonomous' systems that operate in dynamic, unstructured, open environments based on feedback information from a variety of sensors (such as a self-driving car). Regardless of these distinctions, all such systems follow algorithmic instructions that are almost entirely fixed and deterministic, apart from their dependencies on unpredictable sensor data, and narrowly circumscribed probabilistic calculations that are sometimes used for learning and error correction.

6 I use the term 'autonomous weapon system' rather than simply 'autonomous weapon' to indicate that the system may be distributed amongst disparate elements that nonetheless work together to form an autonomous weapon system. For instance, a computer located almost anywhere in the world could receive information from a surveillance drone, and use that information to initiate and direct a strike from a guided weapon system at yet another location, all without human intervention or supervision, thereby constituting an autonomous weapon system. That is, the components of an autonomous weapon system – the sensors, autonomous targeting and decision-making, and the weapon – need not be directly attached to each other or co-located, but merely connected through communications links.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

new wave of technological capability has raised serious concerns and trepidation amongst both the international law community and military professionals as to the moral and legal legitimacy of such systems. As Dr. Jakob Kellenberger, past president of the International Committee of the Red Cross (ICRC), expressed at a conference in San Remo, Italy, in September 2011:

> A truly autonomous system would have artificial intelligence that would have to be capable of implementing IHL. While there is considerable interest and funding for research in this area, such systems have not yet been weaponised. Their development represents a monumental programming challenge that may well prove impossible. The deployment of such systems would reflect a paradigm shift and a major qualitative change in the conduct of hostilities. It would also raise a range of fundamental legal, ethical and societal issues which need to be considered before such systems are developed or deployed. A robot could be programmed to behave more ethically and far more cautiously on the battlefield than a human being. But what if it is technically impossible to reliably program an autonomous weapon system so as to ensure that it functions in accordance with IHL under battlefield conditions? [. . .] [A]pplying pre-existing legal rules to a new technology raises the question of whether the rules are sufficiently clear in light of the technology's specific – and perhaps unprecedented – characteristics, as well as with regard to the foreseeable humanitarian impact it may have. In certain circumstances, states will choose or have chosen to adopt more specific regulations.[7]

As Kellenberger makes clear, there are serious concerns as to whether autonomous technologies will be technically capable of conforming to existing IHL. While many military professionals recognize the technological movement towards greater autonomy in lethal weapons systems, most express strong ethical concerns, including policymakers at the US Office of the Secretary of Defense:

> Restraints on autonomous weapons to ensure ethical engagements are essential, but building autonomous weapons that *fail safely* is the harder task. The wartime environment in which military systems operate is messy and complicated, and autonomous systems must be capable of operating appropriately in it. Enemy adaptation, degraded communications, environmental hazards, civilians in the battlespace, cyber attacks, malfunctions, and 'friction' in war all introduce the possibility that autonomous systems will face unanticipated situations and may act in an unintended fashion. Because they lack a broad contextual intelligence, or common sense, on par with humans, even relatively sophisticated algorithms are subject to failure if they face situations outside their intended design parameters. The complexity of modern computers complicates this problem by making it difficult to anticipate all

7 Jakob Kellenberger, 'Keynote Address', International Humanitarian Law and New Weapon Technologies, 34th Round Table on Current Issues of International Humanitarian Law, San Remo, Italy, 8–10 September 2011, pp. 5–6, available at: http://www.iihl.org/iihl/Documents/JKBSan%20Remo%20Speech.pdf.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

possible glitches or emergent behavior that may occur in a system when it is put into operation.[8]

Because even 'artificially intelligent' autonomous systems must be pre-programmed, and will have only highly limited capabilities for learning and adaptation at best, it will be difficult or impossible to design systems capable of dealing with the fog and friction of war. When we consider the implications of this for protecting civilians in armed conflict, this raises several ethical and legal questions, particularly in relation to conforming to the IHL requirements of the principles of distinction, proportionality, and military necessity, and the difficulty of establishing responsibility and accountability for the use of lethal force.

Autonomous weapon systems raise a host of ethical and social concerns, including issues of asymmetric warfare and risk redistribution from combatants to civilians and the potential to lower the thresholds for nations to start wars.[9] Insofar as such weapons tend to remove the combatants who operate them from area of conflict and reduce the risks of causalities for those who possess them, they tend to also reduce the political costs and risks of going to war. This could result in an overall lowering of the threshold of going to war. Autonomous weapon systems also have the potential to cause regional or global instability and insecurity, to fuel arms races, to proliferate to non-state actors, or initiate the escalation of conflicts outside of human political intentions. Systems capable of initiating lethal force without human supervision could do so even when political and military leadership has not deemed such action appropriate, resulting in the unintended initiation or escalation of conflicts outside of direct human control.[10] Thus, these systems pose a serious threat to international stability and the ability of international bodies to manage conflicts.

In terms of the legal acceptability of these systems under existing IHL,[11] the primary question appears to be whether autonomous systems will be able to satisfy the principles of distinction and proportionality.[12] Given the complexity of these

8 Paul Scharre, 'Why unmanned', in *Joint Force Quarterly*, Issue 61, 2nd Quarter, 2011, p. 92.

9 Peter Asaro, 'How just could a robot war be?', in Adam Briggle, Katinka Waelbers and Philip A. E. Brey (eds), *Current Issues in Computing And Philosophy*, IOS Press, Amsterdam, 2008, pp. 50–64, available at: http://peterasaro.org/writing/Asaro%20Just%20Robot%20War.pdf.

10 By analogy, one should consider the stock market 'Flash Crash' of 6 May 2010, in which automated high-frequency trading systems escalated and accelerated a 1,000-point drop in the Dow Jones average (9%), the single largest drop in history. See *Wikipedia*, 'Flash Crash', available at: http://en.wikipedia.org/wiki/Flash_crash.

11 Noel Sharkey, 'Death strikes from the sky: the calculus of proportionality', in *IEEE Technology and Society Magazine*, Vol. 28, No. 1, 2009, pp. 16–19; Noel Sharkey, 'Saying "no!" to lethal autonomous targeting', in *Journal of Military Ethics*, Vol. 9, No. 4, 2010, pp. 369–383; Markus Wagner, 'Taking humans out of the loop: implications for international humanitarian law', in *Journal of Law Information and Science*, Vol. 21, 2011, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1874039; Matthew Bolton, Thomas Nash and Richard Moyes, 'Ban autonomous armed robots', Article36.org, 5 March 2012, available at: http://www.article36.org/statements/ban-autonomous-armed-robots.

12 See in particular, Articles 51 and 57 of Additional Protocol I to the Geneva Conventions address the protection of the civilian population and precautions in attack. Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, 8 June 1977, 1125 UNTS 3 (entered into force 7 December 1978), available at: http://www2.ohchr.org/english/law/protocol1_2.htm.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

systems, and our inability to foresee how they might act in complex operational environments, unanticipated circumstances, and ambiguous situations, there is a further difficulty – how we can test and verify that a newly designed autonomous weapon system meets the requirements imposed by IHL, as required by Article 36 of Additional Protocol I,[13] and more generally how to govern the increasingly rapid technological innovation of new weapons and tactics.[14]

There is a separate concern that such systems may not have an identifiable operator in the sense that no human individual could be held responsible for the actions of the autonomous weapon system in a given situation, or that the behaviour of the system could be so unpredictable that it would be unfair to hold the operator responsible for what the system does.[15] Such systems might thus eliminate the possibility of establishing any individual criminal responsibility that requires moral agency and a determination of *mens rea*.[16] In the event of an atrocity or tragedy caused by an autonomous weapon system under the supervision or command of a human operator they may also undermine command responsibility and the duty to supervise subordinates, thus shielding their human commanders from what might have otherwise been considered a war crime. It is thus increasingly important to hold states accountable for the design and use of such systems, and to regulate them at an international level.

We are at a juncture at which we must decide how we, as an international community, will treat these systems. Will we treat them as new extensions of old technologies, or as a qualitative shift to a new kind of technology? Is current IHL and IHRL sufficient to deal with autonomous lethal technologies, or are they in need of minor extensions, or major revisions? Is a ban on autonomous weapon systems desirable, or might it disrupt the development of weapons with greater capabilities for respecting moral and legal norms?

It is my view that autonomous weapon systems represent a qualitative shift in military technology, precisely because they eliminate human judgement in the initiation of lethal force. Therefore they threaten to undermine human rights in the absence of human judgement and review. There are good reasons to clarify IHL and IHRL by explicitly codifying a prohibition on the use of autonomous weapon systems. Moreover, these reasons stand up against all of the criticisms offered thus far. The benefits to such a clarification and codification include:

1) avoiding various slippery slopes towards autonomous weapon systems by drawing a principled bound on what can and cannot be automated;

13 The full text of Article 36 of Additional Protocol I on New Weapons reads: 'In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party'.

14 Richard M. O'Meara, 'Contemporary governance architecture regarding robotics technologies: an assessment', in Patrick Lin, Keith Abney and George Bekey, *Robot Ethics*, MIT Press, Cambridge MA, 2011, pp. 159–168.

15 Robert Sparrow, 'Killer robots', in *Journal of Applied Philosophy*, Vol. 24, No. 1, 2007, pp. 62–77.

16 M. Wagner, above note 11, p. 5.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

2) shaping future investments in technology development towards more human-centred designs capable of enhancing ethical and legal conduct in armed conflicts;
3) stemming the potential for more radical destabilizations of the ethical and legal norms governing armed conflict that these new technologies might pose; and
4) establishing the legal principle that automated processes do not satisfy the moral requirements of due consideration when a human life is at stake.

It would therefore be desirable for the international community to move to establish an international ban on autonomous weapon systems on the basis of protecting human rights norms as well as other norms protecting the individual.

Lethal decision-making

In an argument that the use of autonomous weapon systems is morally and legally impermissible, it is necessary to elucidate how autonomous weapon systems fail to meet the necessary and sufficient conditions for permissible killing in armed conflict. It is also necessary to refine the notion of an autonomous weapon system. For now it is sufficient to define the class of autonomous weapon systems as any automated system that can initiate lethal force without the specific, conscious, and deliberate decision of a human operator, controller, or supervisor.

Admittedly, such systems are not unprecedented in the sense that there are various sorts of precursors that have been used in armed conflicts, including mines and other victim-activated traps, as well as certain guided missiles and some automatic defence systems. Indeed, there is a sense in which these systems are not themselves 'weapons' so much as they are automated systems armed with, or in control of, weapons. They thus present a challenge to traditional modes of thought regarding weapons and arms control, which tend to focus on the weapon as a tool or instrument, or upon its destructive effects. Rather, autonomous weapon systems force us to think in terms of 'systems' that might encompass a great variety of configurations of sensors, information processing, and weapons deployment, and to focus on the process by which the use of force is initiated.[17]

Within the US military there has been a policy to follow a human-in-the-loop model when it comes to the initiation of lethal force. The phrase 'human-in-the-loop' comes from the field of human factors engineering, and indicates that a human is an integral part of the system. When it comes to lethal force, the crucial system is the one that contains the decision-making cycle in which any determination to use lethal force is made. In military jargon, this decision cycle is referred to as the 'kill chain', defined in the US Air Force as containing six steps:

17 In the language of Article 36 of Additional Protocol I to the Geneva Conventions, autonomous weapon systems are subject to review on the basis of being a 'new weapon, means or method of warfare'. This implies that using an existing approved weapon in a new way, i.e. with autonomous targeting or firing, is itself subject to review as a new means or method.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

find, fix, track, target, engage and assess.[18] There has been recent discussion of moving to a 'human-on-the-loop' model, in which a human might supervise one or more systems that automate many of the tasks in this six-step cycle. This shift appears to create a middle position between the direct human control of the human-in-the-loop model and an autonomous weapons system. However, the crucial step that determines whether a given system is an autonomous weapon system or not is whether it automates either the target or the engage steps independently of direct human control. We can thus designate the class of systems capable of selecting targets and initiating the use of potentially lethal force without the deliberate and specific consideration of humans as being 'autonomous weapon systems'.

This definition recognizes that the fundamental ethical and legal issue is establishing the causal coupling of automated decision-making to the use of a weapon or lethal force, or conversely the decoupling of human decision-making from directly controlling the initiation of lethal force by an automated system. It is the delegation of the human decision-making responsibilities to an autonomous system designed to take human lives that is the central moral and legal issue.

Note that including a human in the lethal decision process is a necessary, but not a sufficient requirement. A legitimate lethal decision process must also meet requirements that the human decision-maker involved in verifying legitimate targets and initiating lethal force against them be allowed sufficient time to be deliberative, be suitably trained and well informed, and be held accountable and responsible. It might be easy to place a poorly trained person in front of a screen that streams a list of designated targets and requires them to verify the targets, and press a button to authorize engaging those targets with lethal force. Such a person may be no better than an automaton when forced to make decisions rapidly without time to deliberate, or without access to relevant and sufficient information upon which to make a meaningful decision, or when subjected to extreme physical and emotional stress. When evaluating the appropriateness of an individual's decision, we generally take such factors into account, and we are less likely to hold them responsible for decisions made under such circumstances and for any unintended consequences that result, though we do still hold them accountable. Because these factors diminish the responsibility of decision-makers, the design and use of systems that increase the likelihood that decision-making will have to be done under such circumstances is itself irresponsible. I would submit that, when viewed from the perspective of engineering and design ethics, intentionally designing systems that lack responsible and accountable agents is in and of itself unethical, irresponsible, and immoral. When it comes to establishing the standards against which we evaluate lethal decision-making, we should not confuse the considerations we grant to humans acting under duress with our ideals for such standards. Moreover, the fact that we can degrade human performance in such decisions to the level of autonomous systems does not mean we should lower our standards of judging those decisions.

18 Julian C. Cheater, 'Accelerating the kill chain via future unmanned aircraft', Blue Horizons Paper, Center for Strategy and Technology, Air War College, April 2007, p. 5, available at: http://www.au.af.mil/au/awc/awcgate/cst/bh_cheater.pdf.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

While the detailed language defining autonomous weapon systems in an international treaty will necessarily be determined through a process of negotiations, the centrepiece of such a treaty should be the establishment of the principle that human lives cannot be taken without an informed and considered human decision regarding those lives in each and every use of force, and any automated system that fails to meet that principle by removing humans from lethal decision processes is therefore prohibited. This proposal is novel in the field of arms control insofar as it does not focus on a particular weapon, but rather on the manner in which the decision to use that weapon is made. Previous arms control treaties have focused on specific weapons and their effects, or the necessarily indiscriminate nature of a weapon. A ban on autonomous weapons systems must instead focus on the delegation of the authority to initiate lethal force to an automated process not under direct human supervision and discretionary control.

The requirement for human judgement in legal killing

In order for the taking of a human life in armed conflict to be considered legal it must conform to the requirements of IHL. In particular, parties to an armed conflict have a duty to apply the principles of distinction and proportionality. There has been much discussion regarding the ability of autonomous systems to conform to these principles. The most ambitious proposal has been that we may be able to program autonomous weapon systems in such a way that they will conform to the body of IHL, as well as to the specific rules of engagement (ROE) and commander's orders for a given mission.[19] Based in the tradition of constraint-based programming, the proposal is that IHL can be translated into programming rules that strictly determine which actions are prohibited in a given situation. Thus a hypothetical 'ethical governor' could engage to prevent an autonomous weapon system from conducting an action that it determines to be explicitly prohibited under IHL. Arkin further argues that because autonomous weapon systems could choose to sacrifice themselves in situations where we would not expect humans to do the same, these systems might avoid many of the mistakes and failings of humans, and they might accordingly be better at conforming to the rules of IHL than humans.

On its surface, this proposal is quite appealing, and even Kellenberger recognizes its seductive appeal:

> When we discuss these new technologies, let us also look at their possible advantages in contributing to greater protection. Respect for the principles of distinction and proportionality means that certain precautions in attack, provided for in Article 57 of Additional Protocol I, must be taken. This includes the obligation of an attacker to take all feasible precautions in the choice of means and methods of attack with a view to avoiding, and in any event to

19 R. C. Arkin, above note 3, pp. 71–91.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

> minimizing, incidental civilian casualties and damages. In certain cases cyber operations or the deployment of remote-controlled weapons or robots might cause fewer incidental civilian casualties and less incidental civilian damage compared to the use of conventional weapons. Greater precautions might also be feasible in practice, simply because these weapons are deployed from a safe distance, often with time to choose one's target carefully and to choose the moment of attack in order to minimise civilian casualties and damage. It may be argued that in such circumstances this rule would require that a commander consider whether he or she can achieve the same military advantage by using such means and methods of warfare, if practicable.[20]

While it would indeed be advantageous to enhance the protection of civilians and civilian property in future armed conflicts, we must be careful about the inferences we draw from this with regard to permitting the use of autonomous weapon systems. There are a great many assumptions built into this seemingly simple argument, which might mislead us as to the purpose and meaning of IHL.

During armed conflict, the ultimate goal of IHL is to protect those who are not, or are no longer, taking direct part in the hostilities, as well as to restrict the recourse to certain means and methods of warfare. It is tempting to think that this can be objectively and straightforwardly measured. We might like to believe that the principle of distinction is like a sorting rule – that the world consists of civilians and combatants and there is a rule, however complex, that can definitively sort each individual into one category or the other.[21] But it is much more complicated than this. Let's take as an example the difficulty of determining what 'a civilian participating in hostilities' means. The ICRC has laid out a carefully considered set of guidelines for what constitutes 'an act of direct participation in hostilities', and under which a civilian is not afforded the protections normally granted to civilians under IHL.[22] These guidelines set forth three requirements that must be satisfied in order to conclude that a civilian is a legitimate target: 1) threshold of harm, 2) direct causation, and 3) belligerent nexus. Each is elaborated in the ICRC Guidelines, but for present purposes a short summary shall suffice:

> For a specific act to reach the *threshold of harm* required to qualify as direct participation in hostilities, it must be likely to adversely affect the military operations or military capacity of a party to an armed conflict. In the absence of military harm, the threshold can also be reached where an act is likely to inflict death, injury, or destruction on persons or objects protected against direct attack. In both cases, acts reaching the required threshold of harm can only

20 J. Kellenberger, above note 7, p. 6

21 Indeed, there is a tendency in the literature on autonomous weapons to refer to 'discrimination' rather than the principle of distinction, which connotes the 'discrimination task' in cognitive psychology and artificial intelligence. See Noel Sharkey's opinion note in this volume.

22 Nils Mezler, *Interpretive Guidance on the Notion of Direct Participation in Hostilities Under International Humanitarian Law*, ICRC, Geneva, 2009, p. 20, available at: http://www.icrc.org/eng/assets/files/other/icrc-002-0990.pdf.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

> amount to direct participation in hostilities if they additionally satisfy the requirements of direct causation and belligerent nexus. . . .
>
> The requirement of *direct causation* is satisfied if either the specific act in question, or a concrete and coordinated military operation of which that act constitutes an integral part, may reasonably be expected to directly – in one causal step – cause harm that reaches the required threshold. However, even acts meeting the requirements of direct causation and reaching the required threshold of harm can only amount to direct participation in hostilities if they additionally satisfy the third requirement, that of belligerent nexus. . . .
>
> In order to meet the requirement of *belligerent nexus*, an act must be specifically designed to directly cause the required threshold of harm in support of a party to an armed conflict and to the detriment of another. As a general rule, harm caused (A) in individual self-defence or defence of others against violence prohibited under IHL, (B) in exercising power or authority over persons or territory, (C) as part of civil unrest against such authority, or (D) during inter-civilian violence lacks the belligerent nexus required for a qualification as direct participation in hostilities. . . .
>
> Applied in conjunction, the three requirements of *threshold of harm, direct causation* and *belligerent nexus* permit a reliable distinction between activities amounting to direct participation in hostilities and activities which, although occurring in the context of an armed conflict, are not part of the conduct of hostilities and, therefore, do not entail loss of protection against direct attack. Even where a specific act amounts to direct participation in hostilities, however, the kind and degree of force used in response must comply with the rules and principles of IHL and other applicable international law.[23]

These guidelines represent an attempt to articulate a means by which to determine who is a legitimate target and who is not. And yet these are not even called rules – they are called guidelines because they help guide a moral agent through multiple layers of interpretation and judgement. To determine whether a specific individual in a specific circumstance meets each of these requirements requires a sophisticated understanding of a complex situation including: the tactical and strategic implications of a potential harm, as well as the status of other potentially threatened individuals; the nature of causal structures and relations and direct causal implications of someone's actions; the sociocultural and psychological situation in which that individual's intentions and actions qualify as military actions and not, for instance, as the exercise of official duties of authority or personal self-defence.

What does it really mean to say that we can program the rules of IHL into a computer? Is it simply a matter of turning laws written to govern human actions into programmed codes to constrain the actions of machine? Should the next additional protocol to the Geneva Conventions be written directly into computer code? Or is there something more to IHL that cannot be programmed? It is tempting to take an engineering approach to the issue and view the decisions and

23 *Idem*., pp. 50–64.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

INTERNATIONAL REVIEW of the Red Cross

actions of a combatant as a 'black box', and compare the human soldier to the robotic soldier and claim that the one that makes fewer mistakes according to IHL is the 'more ethical' soldier. This has been a common argument strategy in the history of artificial intelligence as well.

There are really two questions here, however. The empirical question is whether a computer, machine, or automated process could make each of these decisions of life and death and achieve some performance that is deemed acceptable. But the moral question is whether a computer, machine or automated process ought to make these decisions of life and death at all. Unless we can prove in principle that a machine should not make such decisions, we are left to wonder if or when some clever programmers might be able to devise a computer system that can do these things, or at least when we will allow machines to make such decisions.

The history of artificial intelligence is instructive here, insofar as it tells us that such problems are, in general, computationally intractable, but if we can very carefully restrict and simplify the problem, we might have better success. We might also, however, compare the sort of problems artificial intelligence has been successful at, such as chess, with the sort of problems encountered in applying IHL requirements. While IHL requirements are in some sense 'rules', they are quite unlike the rules of chess in that they require a great deal of interpretative judgement in order to be applied appropriately in any given situation. Moreover, the context in which the rules are being applied, and the nature and quality of the available information, and alternative competing or conflicting interpretations, might vary widely from day to day, even in the same conflict, or even in the same day.

We might wish to argue that intelligence is uniquely human, but if one can define it specifically enough, or reduce it to a concrete task, then it may be possible to program a computer to do that task better. When we do that, we are necessarily changing the definition of intelligence by redefining a complex skill into the performance of a specific task. Perhaps it is not so important whether we redefine intelligence in light of developments in computing, though it certainly has social and cultural consequences. But when it comes to morality, and the taking of human lives, do we really want to redefine what it means to be moral in order to accommodate autonomous weapon systems? What is at stake if we allow automated systems the authority to decide whether to kill someone? In the absence of human judgement, how can we ensure that such killing is not arbitrary?

Automating the rules of IHL would likely undermine the role they play in regulating ethical human conduct. It would also explain why designers have sought to keep humans-in-the-loop for the purposes of disambiguation and moral evaluation. As Sir Brian Burridge, commander of the British Royal Air Force in Iraq from 2003 to 2005, puts it:

> Under the law of armed conflict, there remains the requirement to assess proportionality and within this, there is an expectation that the human at the end of the delivery chain makes the last assessment by evaluating the situation using rational judgement. Post-modern conflicts confront us … with ambiguous non-linear battlespaces. And thus, we cannot take the human, the

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

> commander, the analyst, those who wrestle with ambiguity, out of the loop. The debate about the human-in-the-loop goes wider than that.[24]

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. Despite the best efforts of its authors to be clear and precise, applying IHL requires multiple levels of interpretation in order to be effective in a given situation. IHL supplements its rules with heuristic guidelines for human agents to follow, explicitly requires combatants to reflexively consider the implications of their actions, and to apply compassion and judgement in an explicit appeal to their humanity. In doing this, the law does not impose a specific calculation, but rather, it imposes a duty on combatants to make a deliberate consideration as to the potential cost in human lives and property of their available courses of action.

Justice cannot be automated

Law is by its essential nature incomplete and subject to interpretation and future review. However careful, thoughtful, and well intentioned a law or rule might be, the legal system is not, and cannot be, perfect. It is a dynamically evolving system, and is designed as such with human institutions to manage its application in the world of human affairs. There are a number of human agents – judges, prosecutors, defenders, witnesses, juries – all of whom engage in complex processes of interpretation and judgement to keep the legal system on track. In short, they are actively engaged in assessing the match between an abstract set of rules and any given concrete situation. The right to due process is essentially the right to have such a deliberative process made publicly accountable.

We could imagine a computer program to replace these human agents, and to automate their decisions. But this, I contend, would fundamentally undermine the right to due process. That right is essentially the right to question the rules and the appropriateness of their application in a given circumstance, and to make an appeal to informed human rationality and understanding. Do humans in these positions sometimes make mistakes? Yes, of course they do. Human understanding, rationality, and judgement exceed any conceivable system of fixed rules or any computational system, however. Moreover, when considering the arguments in a given case, the potential for appeals to overturn judicial decisions, and the ways in which opinions and case law inform the interpretation of laws, we must recognize that making legal judgements requires considering different, incompatible, and even contradictory perspectives, and drawing insight from them. There are no known computational or algorithmic systems that can do this, and it might well be impossible for them to do so.

24 Brian Burridge, 'UAVs and the dawn of post-modern warfare: a perspective on recent operations', in *RUSI Journal*, Vol. 148, No. 5, October 2003, pp. 18–23.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

INTERNATIONAL REVIEW of the Red Cross

More importantly, human judgement is constitutive of the system of justice. That is, if any system of justice is to apply to humans, then it must rely upon human reason. Justice itself cannot be delegated to automated processes. While the automation of various tasks involved in administrative and legal proceedings may enhance the ability or efficiency of humans to make their judgements, it cannot abrogate their duty to consider the evidence, deliberate alternative interpretations, and reach an informed opinion. Most efforts at automating administrative justice have not improved upon human performance, in fact, but have greatly degraded it.[25] To automate these essential aspects of human judgement in the judicial process would be to dehumanize justice, and ought to be rejected in principle.

In saying that the automation of human reasoning in the processes of justice ought to be rejected in principle, I mean that there is no automated system, and no measure of performance that such a system could reach, that we should accept as a replacement for a human. In short, when it comes to a system of justice, or the state, or their agents, making determinations regarding the human rights of an individual, the ultimate agents and officials of the state must themselves be human. One could argue for this principle on moral grounds, as well as on the legal grounds that it is constitutive of, and essential to, the system of justice itself independently of its moral standing.

Within the military there are many layers of delegated authority, from the commander-in-chief down to the private, but at each layer there is a responsible human to bear both the authority and responsibility for the use of force. The nature of command responsibility does not allow one to abdicate one's moral and legal obligations to determine that the use of force is appropriate in a given situation. One might transfer this obligation to another responsible human agent, but one then has a duty to oversee the conduct of that subordinate agent. Insofar as autonomous weapon systems are not responsible human agents, one cannot delegate this authority to them.

In this sense, the principle of distinction can be seen not simply as following a rule that sorts out combatants from civilians, but also of giving consideration to human lives that might be lost if lethal force is used. And in this regard, it is necessary for a human being to make an informed decision before that life can be taken. This is more obvious 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.

This is why we cannot claim that an autonomous weapon system would be morally superior to a human soldier on the basis that it might be technologically capable of making fewer errors in a discrimination task, or finding means of neutralizing military targets that optimally minimize the risk of disproportionate harms. This is not to say that these goals are not desirable. If technologies did exist

25 Danielle Keats Citron, 'Technological due process', in *Washington University Law Review*, Vol. 85, 2008, pp. 1249–1292.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

that could distinguish civilians from combatants better than any human, or better than the average combatant, then those technologies should be deployed in a manner to assist the combatant in applying the principle of distinction, rather than used to eliminate human judgement. Similarly, if a technology were capable of determining a course of action which could achieve a military objective with minimal collateral damage, and minimize any disproportionate harms, then that technology could be employed by a human combatant charged with the duty of making an informed choice to initiate the use of lethal force in that situation.

Any automated process, however good it might be, and even if measurably better than human performance, ought to be subject to human review before it can legitimately initiate the use of lethal force. This is clearly technologically required for the foreseeable future because autonomous systems will not reach human levels of performance for some time to come. But more importantly, this is a moral requirement and, in many important instances, a legal requirement. I therefore assert that in general there is a duty not to permit autonomous systems to initiate lethal force without direct human supervision and control.

There are two basic strategies for arguing that autonomous weapons systems might provide a morally or legally superior means of waging war compared to current means of armed conflict. There are many variations of the argument, which I divide into two classes: 1) pragmatic arguments pointing to failures of lethal decision-making in armed conflict and arguing to possible/hypothetical technological improvements through automating these decisions,[26] and 2) arguing that insofar as such systems imply a reduced risk to combatants and/or civilians in general, as measured by fewer casualties, there is a moral imperative to use them. Such arguments have been made for precision weapons in the past,[27] and more recently for Predator drones and remote-operated lethality.[28]

Are more precise weapons more 'moral' than less precise weapons? It is easy enough to argue that given the choice between attacking a military target with a precision-guided munition with low risk of collateral damage, and attacking the same target by carpet bombing with a high risk or certainty of great collateral damage, one ought to choose the precision-guided munition. That is the moral and legal choice to make, all other things being equal. Of course, there is quite a bit that might be packed into the phrase 'all other things being equal'. Thus it is true that one should prefer a more precise weapon to a less precise weapon when deciding how to engage a target, but the weapon is not ethically independent of that choice. And ultimately it is the human agent who chooses to use the weapon that is judged to be moral or not. Even the most precise weapon can be used illegally and immorally. All that precision affords is a possibility for more ethical behaviour – it does not determine or guarantee it.

26 Ronald C. Arkin, 'Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture', Georgia Institute of Technology, Technical Report GUT-GVU-07-11, 2007, p. 11.

27 Human Rights Watch, 'International humanitarian law issues in the possible U.S. invasion of Iraq', in *Lancet*, 20 February 2003.

28 Bradley Jay Strawser, 'Moral predators: the duty to employ uninhabited aerial vehicles', in *Journal of Military Ethics*, Vol. 9, No. 4, 2010, pp. 342–368.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

This may seem like a semantic argument, but it is a crucial distinction. We do not abrogate our moral responsibilities by using more precise technologies. But as with other automated systems, such as cruise control or autopilot, we still hold the operator responsible for the system they are operating, the ultimate decision to engage the automated system or to disengage it, and the appropriateness of these choices. Indeed, in most cases these technologies, as we have seen in the use of precision-guided munitions and armed drones, actually increase our moral burden to ensure that targets are properly selected and civilians are spared. And indeed, as our technologies increase in sophistication, we should design them so as to enhance our moral conduct.

There is something profoundly odd about claiming to improve the morality of warfare by automating humans out of it altogether, or at least by automating the decisions to use lethal force. The rhetorical strategy of these arguments is to point out the moral shortcomings of humans in war – acts of desperation and fear, mistakes made under stress, duress, and in the fog of war. The next move is to appeal to a technological solution that might eliminate such mistakes. This might sound appealing, despite the fact that the technology does not exist. It also misses two crucial points about the new kinds of automated technologies that we are seeing. First, that by removing soldiers from the immediate risks of war, which tele-operated systems do without automating lethal decisions, we can also avoid many of these psychological pressures and the mistakes they cause. Second, if there were an automated system that could outperform humans in discrimination tasks, or proportionality calculations, it could just as easily be used as an advisory system to assist and inform human decision-makers, and need not be given the authority to initiate lethal force independently of informed human decisions.[29]

Arguments against banning autonomous weapon systems

In a recent policy brief, Anderson and Waxman offer a criticism of proposals to ban autonomous weapon systems.[30] They conclude that while it is important to establish international norms regarding the use of autonomous weapon systems, a ban is not the best way to do it. There are, however, numerous problems with their argument and many of their conclusions. The main thrust of their argument is based in two assumptions:

> Recognizing the inevitable but incremental evolution of these technologies is key to addressing the legal and ethical dilemmas associated with them; US policy toward resolving those dilemmas should be built upon these assumptions. The certain yet gradual development and deployment of these systems, as well as the humanitarian advantages created by the precision of some systems,

29 Peter Asaro, 'Modeling the moral user: designing ethical interfaces for tele-operation', in *IEEE Technology & Society*, Vol. 28, No. 1, 2009, pp. 20–24, available at: http://peterasaro.org/writing/Asaro%20Modeling%20Moral%20User.pdf.

30 K. Anderson and M. C. Waxman, above note 3, p. 13.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

> make some proposed responses – such as prohibitory treaties – unworkable as well as ethically questionable.[31]

Here we see several arguments being made against the proposal for an international prohibitory treaty. First, they insist upon starting from the assumptions that these technologies are inevitable, and that their development is incremental. Yet they provide no evidence or arguments to support either assumption, even though there are strong reasons to reject them. They then make a further argument that some of these systems may have humanitarian advantages, and thus prohibitions are both 'unworkable' and 'ethically questionable'. Having just explained why it is not 'ethically questionable' to argue that even the most precise autonomous weapon systems jeopardize human rights, I want to focus on their two preliminary assumptions and what they might mean for the practicality of an international prohibition.

Are autonomous weapon systems inevitable?

Why should we assume that autonomous weapon systems are inevitable? What might this actually mean? As a philosopher and historian of science and technology, I often encounter claims about the 'inevitability' of scientific discoveries or technological innovations. The popularity of such claims is largely due to the retrospective character of history, and applying our understanding of past technologies to thinking about the future. That is, it seems easy for us, looking back, to say that the invention of the light bulb, or the telephone, or whatever technology you prefer was inevitable – because it did in fact happen. It is hard to imagine what the world would be like if it had not happened. Yet when one looks carefully at the historical details, whether a technology succeeded technologically was in most cases highly contingent on a variety of factors. In most cases, the adoption of the technology was not guaranteed by the success of the innovation, and the means and manner of its eventual use always depended upon a great variety of social and cultural forces. Indeed, when we look at the great many technological failures, and indeed the many failed attempts to commercialize the light bulb before it finally succeeded, what becomes clear is that very few, if any, technologies can fairly be claimed to be 'inevitable'. And even the successful light bulb was dependent upon the innovation and development of electrical utilities, and a host of other electric appliances, such as toasters, for its widespread adoption. Technologies evolve much faster now, but they are just as dynamic and unpredictable.

Perhaps what Anderson and Waxman mean is that it seems very likely that these technologies will be developed. This seems more plausible. Indeed, simplistic systems can already implement the essential elements of an autonomous weapon system, though these would fail to meet the existing international legal standards of discrimination and proportionality.[32] But even ignoring the existing legal

31 *Idem*., p. 2.
32 M. Wagner, above note 11, pp. 5–9.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

limitations, the fact that we can build autonomous lethal technologies does not mean we will use them. Given that various sorts of autonomous weapon systems are already possible, it might be claimed that it is their adoption that is inevitable. But to assume this would be glossing over the important differences between the invention of a technology and its widespread adoption in society. There are certainly some strong motivations for adopting such technologies, including the desire to reduce the risks to military personnel, as well as reduce the costs and number of people needed for various military operations and capabilities.

Or more strongly, Anderson and Waxman might mean that we should assume that it is inevitable that there will be autonomous weapon systems that are capable of meeting the requirements of some measure of discrimination and proportionality. But this is an empirical claim, about the capabilities of technologies that do not yet exist, being measured against a metric that does not yet exist. As a purely empirical question, these technologies may or may not come into existence and we may not even be able to agree upon acceptable metrics for evaluating their performance, so why should we believe that they are inevitable?[33]

The crucial question here is whether these technologies can meet the requirements of international law, and this is far from certain. The arguments claiming the ethical superiority of robotic soldiers sound suspiciously like claims from the early days of artificial intelligence that computers would someday beat human grandmasters at chess. And, forty years later than initial predictions, IBM's Deep Blue did manage to beat Gary Kasparov. But there are important differences between chess and IHL that are worth noting. Chess is a fairly well-defined rule-based game that is susceptible to computational analysis. Ultimately, the game of chess is not a matter of interpretation, nor is it a matter of social norms. International law, while it has rules, is not like chess. Law always requires interpretation and judgement in order to apply it to real world situations. These interpretations and judgements are aided by historical precedents and established standards, but they are not strictly determined by them. The body of case law, procedures, arguments, and appeals is able to defend old principles or establish new precedents, and thereby establish norms and principles, even as those norms and principles continue to grow in meaning over time.

Thus, insisting that autonomous weapon systems are inevitable is actually quite pernicious. On the one hand, this assumption would make the establishment of a ban seem automatically impractical or unworkable. That is, if we start from the assumption that the banned systems will exist and will be used, then why should we bother to ban them? But of course, they do not exist and are not being used, and even if they were being used already they could still be banned going forward. And far from being unworkable or impractical, a ban could be quite effective in shifting innovation trajectories towards more useful and genuinely ethical systems. It seems

33 For comparison, consider electric cars, a technology that has existed for a century. Even with the recent popularity of hybrid gas/electric cars, and some highly capable electric cars, few people would endorse the claim that our transition to electric cars is inevitable. And this is a technology that is already possible, i.e. it exists.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

straightforward that we can define the class of autonomous weapon systems clearly enough, and then debate how a treaty might apply to, or exempt, certain borderline cases such as reactive armour, anti-ballistic missile defences, or supervisory systems. A ban cannot be expected to prohibit each and every use of automation in armed conflict, but rather to establish an international norm that says that it is illegitimate to use systems that make automated lethal decisions. The international bans on landmines and cluster munitions may have not completely eliminated landmines and cluster munitions and their use in armed conflict, but they have made it more difficult for manufacturers to produce them profitably, and for militaries to use them without repercussions in the international community.

Moreover, starting from an assumption of the inevitability of autonomous weapon systems appears to make the acceptability of such systems a foregone conclusion. Yet what is ultimately at issue here is what the international standards of acceptability will be – what the international community will consider the norms of conduct to be. To assume the inevitability of the development and use of the technologies in question is to close off further discussion on the wisdom and desirability of pursuing, developing, and using these technologies. In short, the development and use of autonomous weapon systems is not inevitable – no technology is. Yes, they are possible; if they were not then there would be no need to ban them, but their developments still requires great investment. And even if we may not be able to prevent the creation of certain technologies, we will always be able to assert a position on the moral and legal acceptability of their use. It does not follow that simply because a technology exists, its use is acceptable.

So what if autonomous weapon systems are developing incrementally?

I want to return to the second assumption insisted upon by Anderson and Waxman, namely that autonomous weapons systems will develop 'incrementally'. What is this assumption meant to do for their argument? Again, from the perspective of technological development, all technologies develop incrementally in some sense. Why would this change the way in which we address their ethical and legal implications? Perhaps Anderson and Waxman are merely trying to disarm the fear that soldiers will be replaced by robots in some great technological leap. As their argument continues, it becomes clearer that what they have in mind is that the transition to autonomous weapon systems will happen in incremental 'baby steps', each of which will be carefully considered and scrutinized. This is a rather inventive inversion of the slippery slope argument. Instead of asserting that these technologies are dangerous because they encourage us to delegate more and more authority to automated systems, eventually resulting in automated systems with the illegitimate but *de facto* authority to target and kill humans, it argues that such systems will be legitimate because each step along the way seems acceptable. They appear to argue that we should accept the end result of this line of reasoning because we were able to reach it through a series of moral adjustments, none of which on its own seemed too awful.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

It would make more sense to see this as a slippery slope leading us to a result we believe to be unacceptable. This should lead us to look more carefully for an underlying principle upon which we can stop the perilous slide to an undesirable conclusion. And indeed, there is a principled boundary that we can establish with regard to autonomous weapon systems. That boundary is that for any system capable of initiating lethal force, a human being needs to be meaningfully involved in making the decision of whether or not lethal force will actually be used in each case. And while we can blur this line with various technological systems of shared and supervisory control, we could also design those systems in such ways as to make this line clearer, and to make those decisions better informed.[34]

On the need to establish norms

The conclusions that Anderson and Waxman draw are wrong about the implications of a ban on autonomous weapon systems, but they are correct about the significance of the establishment of norms regarding their use and the need for some constraints:

> The United States must act, however, before international expectations about these technologies harden around the views of those who would impose unrealistic, ineffective or dangerous prohibitions or those who would prefer few or no constraints at all.[35]

What they recognize is that there is a new moral space being opened up by these technologies, and it is not yet settled what the international community will accept as the new norms of warfare in the age of robotics and automation. They are also correct that the US, as both a super power and the leader in developing many of these new technologies, is in a unique position to establish the precedents and norms that will shape the future of armed conflict. More obviously, Anderson and Waxman have not shown how banning autonomous weapon systems is unrealistic, nor have they shown any evidence that such a ban would be ineffective or immoral. Let us consider each of these claims in turn.

How might we make sense of the claim that a ban on autonomous weapon systems would be unrealistic? Is it that such a ban would, in practice, be difficult to implement? All arms control treaties pose challenges in their implementation, and a ban on autonomous weapon systems should not prove exceptionally more or less difficult than others, and therefore is not unrealistic in this sense. Or is the claim that it would be politically difficult to find support for such a ban? In my personal experience, there are a great many individuals, particularly among military officers and policymakers but also among engineers and executives in the defence industry, who would support such a ban. Moreover, it is clear, from my experiences in engaging with the public, that strong moral apprehensions about automated weapons systems are broad-based, as is fear of the potential risks they pose. At the

34 P. Asaro, above note 29, pp. 20–24.
35 K. Anderson and M. C. Waxman, above note 3, p. 2.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

very least, a ban is not unrealistic in the sense that it might likely find broad public and official support.

Indeed, the only way we might consider such a ban to be unrealistic is if we accept Anderson and Waxman's unwarranted assumption that these systems are inevitable. If we accept that as a foregone conclusion, then attempting to halt the inevitable does seem unrealistic. But there is nothing inevitable about an emerging technology, the capabilities of which do not yet exist, and the norms surrounding which have not yet been established.

Anderson and Waxman also anticipate an objection to autonomous weapon systems on the basis of a moral principle:

> A second objection is a moral one, and says that it is simply wrong per se to take the human moral agent entirely out of the firing loop. A machine, no matter how good, cannot completely replace the presence of a true moral agent in the form of a human being possessed of a conscience and the faculty of moral judgement (even if flawed in human ways). In that regard, the title of this essay is deliberately provocative in pairing 'robot' and 'soldier', because, on this objection, that is precisely what should never be attempted.
>
> This is a difficult argument to address, since it stops with a moral principle that one either accepts or does not.[36]

The objection they refer to draws upon what is supposed to be a sort of stand-alone principle.[37] Therefore, they suppose that there is no justification for accepting it apart from one's own moral intuitions. I would submit that the arguments presented above in this article have demonstrated that the moral principle for rejecting autonomous weapon systems is in fact implicit within IHL through its various anthropocentric formulations and requirements. Moreover, it is implicit within the very structure of law, the processes of justice, and due process in particular. We require the presence of a human as a legal agent, independent of the moral requirement that they be moral agents.

It is not simply that the decision to kill is a weighty one, though it is. The decision to kill a human can only be legitimate if it is non-arbitrary, and there is no way to guarantee that the use of force is not arbitrary without human control, supervision, and responsibility. It is thus immoral to kill without the involvement of human reason, judgement, and compassion, and it should be illegal.

Conclusion

As a matter of the preservation of human morality, dignity, justice, and law we cannot accept an automated system making the decision to take a human life. And we should respect this by prohibiting autonomous weapon systems. When it comes

36 *Idem*., p. 11.
37 M. Bolton, T. Nash and R. Moyes, above note 11.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768

to killing, each instance is deserving of human attention and consideration in light of the moral weight inherent in the active taking of a human life.

As technology advances, it gives humanity greater control over the world. With that new control comes increased responsibility. While this seems obvious for technologies that influence human and environmental welfare, it is also true for military technologies. While the development of advanced military technologies does not necessarily imply that they will be, or can be, used more carefully and ethically, that possibility exists. But new capabilities also bring with them a potential to regress in ethics and morality, rather than progress. Ultimately the nature of our moral progress in the conduct of war depends upon our technology in a deeper sense than merely enabling combatants to conduct wars with fewer casualties, and goes beyond the requirements of IHL and IHRL. In choosing the weapons and tactics with which we engage in armed conflict, we are also making a moral choice about the world we wish to live in and fight for, and the legitimate conditions under which we can bring that world into being. In making such choices, we must resist arguments that any end is either so desirable or undesirable that any means of achieving it are acceptable. We must also acknowledge that the means by which we enact change in the world, or resist change, thereby become an aspect of that world. If we truly wish to build a future in which armed conflict is both unnecessary and unacceptable, we must arrive there through a process that raises our moral standards with each new technological innovation, rather than by lowering those standards.

The international community should begin discussions on the formation of a treaty to ban autonomous weapons systems. Insofar as such systems do not yet exist, such a ban would help to focus the development of future military technologies away from these so-called ethical systems and towards the development of systems that can actually improve the ethical conduct of humans in armed conflicts. The critics of such a ban base their criticisms on unsupported claims about the inevitability of these technologies and misleading claims about ethically enhanced technologies. For as long as their potential capabilities remain uncertain, these technologies are emerging in a dynamic field of ethical and legal norms. Though we might like to trust in the promise of more ethical wars through these hypothetical autonomous weapon systems, the reality is that they can also degrade our conceptions and standards of ethical conduct, and distract us from developing the technological enhancement of human moral reasoning by pursuing an improbable technology that threatens to undermine our human rights on a fundamental level. It might also distract us from enhancing and improving IHL and IHRL to deal appropriately and morally with these new technologies.

Downloaded from https://www.cambridge.org/core. Tsinghua University, on 31 Jan 2018 at 07:31:24, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1816383112000768