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

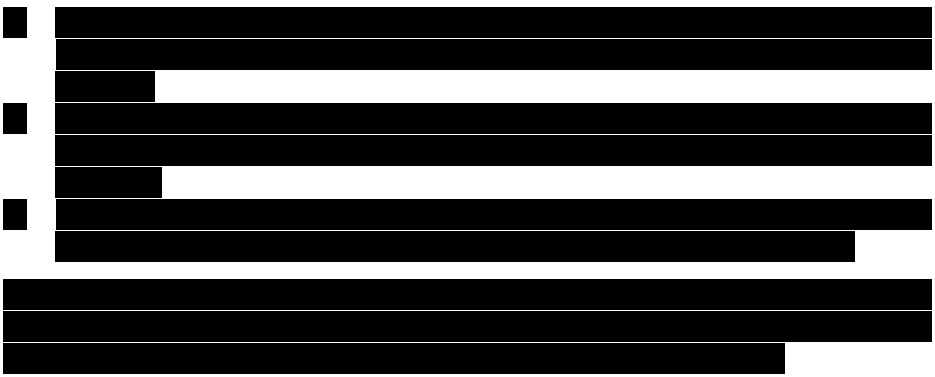
Peter Asaro

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

doi:10.1017/S1816383112000768



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.

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.

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.

> 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.

> 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.

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

> 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.

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.

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.

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]