Volume 94 Number 886 Summer 2012

INTERNATIONAL REVIEW of the Red Cross

COMMENTS AND OPINIONS

The evitability of autonomous robot warfare*

Noel E. Sharkey**

Noel E. Sharkey is Professor of Artificial Intelligence and Robotics and Professor of Public Engagement in the Department of Computer Science at the University of Sheffield, UK, and currently holds a Leverhulme Research Fellowship on an ethical and technical assessment of battlefield robots.

Abstract

This is a call for the prohibition of autonomous lethal targeting by free-ranging robots. This article will first point out the three main international humanitarian law (IHL)/ ethical issues with armed autonomous robots and then move on to discuss a major stumbling block to their evitability: misunderstandings about the limitations of robotic systems and artificial intelligence. This is partly due to a mythical narrative from science fiction and the media, but the real danger is in the language being used by military researchers and others to describe robots and what they can do. The article will look at some anthropomorphic ways that robots have been discussed by the military and then go on to provide a robotics case study in which the language used obfuscates the IHL issues. Finally, the article will look at problems with some of the current legal instruments and suggest a way forward to prohibition.

Keywords: autonomous robot warfare, armed autonomous robots, lethal autonomy, artificial intelligence, international humanitarian law.

* The title is an allusion to a short story by Isaac Asimov, 'The evitable conflict', where 'evitable' means capable of being avoided. Evitability means avoidability.

** Thanks for comments on earlier drafts go to Colin Allen, Juergen Altmann, Niall Griffith, Mark Gubrud, Patrick Lin, George Lucas, Illah Nourbakhsh, Amanda Sharkey, Wendell Wallach, Alan Winfield, and to editor-in-chief Vincent Bernard and the team of the *International Review of the Red Cross*, as well as others who prefer to remain unnamed.

We could be moving into the final stages of the industrialization of warfare towards a factory of death and clean-killing where hi-tech countries fight wars without risk to their own forces. We have already seen the exponential rise of the use of drones in the conflicts in Iraq and Afghanistan and by the US Central Intelligence Agency for targeted killings and signature strikes in countries outside the war zones: Pakistan, Yemen, Somalia, and the Philippines. Now more than fifty states have acquired or are developing military robotics technology.[1]

All of the armed robots currently in use have a person in the loop to control their flight and to apply lethal force. But that is set to change soon. Over the last decade the roadmaps and plans of all US forces have made clear the desire and intention to develop and use autonomous battlefield robots. Fulfilment of these plans to take the human out of the control loop is well underway for aerial, ground, and underwater vehicles. And the US is not the only country with autonomous robots in their sights. China, Russia, Israel, and the UK are following suit. The end goal is a network of land, sea, and aerial robots that will operate together autonomously to locate their targets and destroy them without human intervention.[2]

IHL and ethical issues with lethal autonomous robots

A major IHL issue is that autonomous armed robot systems cannot discriminate between combatants and non-combatants or other immune actors such as service workers, retirees, and combatants that are wounded, have surrendered, or are mentally ill in a way that would satisfy the principle of distinction. There are systems that have a weak form of discrimination. For example, the Israeli Harpy is a loitering munition that detects radar signals. When it finds one, it looks at its database to find out if it is friendly and if not, it dive bombs the radar. This type of discrimination is different from the requirements of the principle of distinction because, for example, the Harpy cannot tell if the radar is on an anti-aircraft station or on the roof of a school.

Robots lack three of the main components required to ensure compliance with the principle of distinction. First, they do not have adequate sensory or vision processing systems for separating combatants from civilians, particularly in insurgent warfare, or for recognizing wounded or surrendering combatants. All that is available to robots are sensors such as cameras, infrared sensors, sonars, lasers, temperature sensors, and ladars etc. These may be able to tell us that something is a human, but they could not tell us much else. There are systems in the labs that can recognize still faces and they could eventually be deployed for individual targeting in limited circumstance. But how useful could they be with

1 Noel Sharkey, 'The automation and proliferation of military drones and the protection of civilians', in *Journal of Law, Innovation and Technology*, Vol. 3, No. 2, 2001, pp. 229–240.

2 Noel Sharkey, 'Cassandra or the false prophet of doom: AI robots and war', in *IEEE Intelligent Systems*, Vol. 23, No. 4, 2008, pp. 14–17.

moving targets in the fog of war or from the air? British teenagers beat the surveillance cameras simply by wearing hooded jackets.

Second, a computer can compute any given procedure that can be written down in a programming language. This is rather like writing a knitting pattern or recipe. We also need to be able to specify every element in sufficient detail for a computer to be able to operate on it. The problem for the principle of distinction is that we do not have an adequate definition of a civilian that we can translate into computer code. The laws of war does not provide a definition that could give a machine with the necessary information. The 1949 Geneva Convention requires the use of common sense, while the 1977 Protocol I defines a civilian in the negative sense as someone who is not a combatant.[3]

Third, even if machines had adequate sensing mechanisms to detect the difference between civilians and uniform-wearing military, they would still be missing battlefield awareness or common sense reasoning to assist in discrimination decisions. We may move towards having some limited sensory and visual discrimination in certain narrowly constrained circumstances within the next fifty years. However, I suspect that human-level discrimination with adequate common sense reasoning and battlefield awareness may be computationally intractable.[4] At this point we cannot rely on machines ever having the independent facility to operate on the principle of distinction as well as human soldiers can.[5] There is no evidence or research results to suggest otherwise.

A second IHL issue is that robots do not have the situational awareness or agency to make proportionality decisions. One robotics expert has argued that robots could calculate proportionality better than humans.[6] However, this concerns the *easy proportionality problem*: minimizing collateral damage by choosing the most appropriate weapon or munition and directing it appropriately. There is already software called bugsplat used by the US military for this purpose. The problem is that it can only ease collateral impact. For example, if munitions were used near a local school where there were 200 children, the appropriate software may mean that only fifty children were killed rather than all had a different bomb been used.

The *hard proportionality problem* is making the decision about whether to apply lethal or kinetic force in a particular context in the first place. What is the balance between loss of civilian lives and expected military advantage? Will a particular kinetic strike benefit the military objectives or hinder them because it upsets the local population? The list of questions is endless. The decision about what is proportional to direct military advantage is a human qualitative and subjective

3 Article 50(1) of the Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts, 8 June 1977 (hereinafter Additional Protocol I)

4 As a scientist I cannot exclude the notion that some black swan event could change my scepticism, but at present we certainly cannot rely on this as a credible option in discussions of lethal force and the protection of innocents.

5 See Noel E. Sharkey, 'Grounds for Discrimination: Autonomous Robot Weapons', in *RUSI Defence Systems*, Vol. 11, No. 2, 2008, pp. 86–89.

6 Ronald C. Arkin, *Governing Lethal Behavior in Autonomous Systems*, CRC Press Taylor & Francis Group, Boca Raton F.L., 2009, pp. 47–48.

decision. It is imperative that such decisions are made by responsible, accountable human commanders who can weigh the options based on experience and situational awareness. When a machine goes wrong it can go really wrong in a way that no human ever would.

I turn to the well-known expression that much of war is art and not science. Or as Col. David M. Sullivan, an Air Force pilot with extensive experience with both traditional and drone airstrikes from Kosovo to Afghanistan, told *Discover* magazine: 'If I were going to speak to the robotics and artificial intelligence people, I would ask, "How will they build software to scratch that gut instinct or sixth sense?" Combat is not black-and-white'.[7]

Arguing against a ban on lethal autonomous robot weapons, Anderson and Waxman state that some leading roboticists have been working on creating algorithms to capture the two fundamental principles of distinction and proportionality. But they cite only one roboticist: 'One of the most ambitious of these efforts is by roboticist Ronald C. Arkin, who describes his work on both distinction and proportionality in his "Governing Lethal Behavior".'[8] But this is mistaken because while this roboticist discusses both principles, he is not conducting research on either of them. He only suggests that they will be solvable by machines one day.

The work Anderson and Waxman cite is, in fact, merely a suggestion for a computer software system for the ethical governance of robot 'behaviour'.[9] This is what is known as a 'back-end system'. Its operation relies entirely on information from systems yet 'to be developed' by others sometime in the future. It has no direct access to the real world through sensors or a vision system and it has no means to discriminate between combatant and non-combatant, between a baby and a wounded soldier, or a granny in a wheelchair and a tank. It has no inference engine and certainly cannot negotiate the types of common sense reasoning and battlefield awareness necessary for discrimination or proportionality decisions. There is neither a method for interpreting how the precepts of the laws of war apply in particular contexts nor is there any method for resolving the ambiguities of conflicting laws in novel situations.

A third issue is accountability.[10] A robot does not have agency, moral or otherwise, and consequently cannot be held accountable for its actions. Moreover, if autonomous robots were used in limited circumstances in the belief that they could operate with discrimination, it would be difficult to decide exactly who was accountable for mishaps. Some would say that the commander who gave the order to send the robot on a mission would be responsible (last point of contact). But that would not be fair since it could be the fault of the person who programmed the

7 Mark Anderson, 'How Does a Terminator Know When to Not Terminate', in *Discover Magazine*, May 2010, p. 40.

8 Kenneth Anderson and Matthew Waxman, 'Law and Ethics of Robot Soldiers', in *Policy Review*, in press 2012.

9 See R. C. Arkin, above note 6.

10 Robert Sparrow, 'Building a Better WarBot: Ethical Issues in the Design of Unmanned Systems for Military Applications', in *Science and Engineering Ethics*, Vol. 15, No. 2, 2009, pp.169–187.

INTERNATIONAL REVIEW of the Red Cross

mission, the manufacturer who made the robot, or the senior staff or policymakers who decided to deploy it. Or it could be claimed that the device was tampered with or damaged. Anderson and Waxman dismiss the accountability objection out of hand:

> Post hoc judicial accountability in war is just one of many mechanisms for promoting and enforcing compliance with the laws of war, and devotion to individual criminal liability as the presumptive mechanism of accountability risks blocking development of machine systems that would, if successful, reduce actual harms to civilians on or near the battlefield.[11]

But I disagree. Using a weapon without a clear chain of accountability is not a moral option. Without accountability to enforce compliance many more civilian lives could be endangered.

On the basis of these three issues, I will argue here that the morally correct course of action is to ban autonomous lethal targeting by robots. Before looking at problems with the legal instruments, I will first examine a major stumbling block to a prohibition on the development of armed autonomous robots. A notion proposed by the proponents of lethal autonomous robots is that there are technological 'fixes' that will make them behave more ethically and more humanely than soldiers on the battlefield. I will argue here that this has more to do with descriptive language being used to describe robots rather than what robots can actually do.

Anthropomorphism and mythical artificial intelligence

The common conception of artificial intelligence (AI) and robotics has been distorted by the cultural myth of AI engendered partly by science fiction, by media reporting, and by robotics experts sucked into the myths or seeking public recognition. Robots can be depicted as sentient machines that can think and act in ways superior to humans and that can feel emotions and desires. This plays upon our natural tendency to attribute human or animal properties and mental states (anthropomorphism or zoomorphism) to inanimate objects that move in animal-like ways.[12] We are all susceptible to it and it is what has made puppets so appealing to humans since ancient times.

The myth of AI makes it acceptable, and even customary, to describe robots with an anthropomorphic narrative. Journalists are caught up in it and know that their readers love it. But we cannot just blame the media. It is a compelling narrative and even some roboticists inadvertently feed into the myth. Like other cultural myths, it can be harmless in casual conversations in the lab. But it is a perilous road to follow in legal and political discussions about enabling machines to apply lethal force.

11 See K. Anderson and M. Waxman, above note 8.

12 Amanda Sharkey and Noel Sharkey, 'Artificial Intelligence and Natural Magic', in *Artificial Intelligence Review*, Vol. 25, No. 1–2, 2006, pp. 9–19.

Even with remote-controlled robots, anthropomorphism catches the military. The *Washington Post* reported that soldiers on the battlefield using bomb disposal robots often treat them as fellow warriors and are sometimes prepared to risk their own lives to save them. They even take them fishing during leisure time and get them to hold a fishing rod in their gripper.[13] In the mid-1990s, roboticist Mark Tilden ran a test of his 'Bagman' multipede mine-clearing robot at the Yuma Arizona Missile testing range. Each time that the robot detected a mine, it stamped on it and one leg was blown off. A US colonel watching the legs being blown off one by one finally called a halt to the test because he felt that it was inhumane.[14]

The impact of anthropomorphism can go all the way to the top. Gordon Johnson, former head of the Joint Forces Command at the Pentagon, told the *New York Times* that robots 'don't get hungry. They're not afraid. They don't forget their orders. They don't care if the guy next to them has just been shot.'[15] All of this can also be said of a landmine and my washing machine. Yet if Johnson had said it about these devices, it would have sounded ridiculous. Without being directly anthropomorphic, Johnson is leaking it.

Similarly, Marchant et al. say of robots that 'they can be designed without emotions that cloud their judgment or result in anger and frustration with ongoing battlefield events'.[16] This leaks anthropomorphism because it implies that without special design the robots would have emotions to cloud their judgements. Clearly this is wrong. The myth of robot soldiers even spreads into the law community with titles like 'Law and Ethics of Robot Soldiers'.[17]

A case study of wishful mnemonics

In his influential paper, 'Artificial intelligence meets natural stupidity',[18] Drew McDermott, a Professor of AI at Yale University, expressed concern that the discipline of AI could ultimately be discredited by researchers using natural language mnemonics, such as 'UNDERSTAND', to describe aspects of their programs. Such terms describe a researcher's aspirations rather than what the programs actually do. McDermott called such aspirational terms 'Wishful Mnemonics' and suggested that, in using them, the researcher 'may mislead a lot of people, most prominently himself', that is, the researcher may misattribute

13 Joel Garreau, 'Bots on the Ground', in *Washington Post*, 6 May 2007, available at: http://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009.html (last visited January 2012).

14 Mark Tilden, personal communication and briefly reported in *ibid.*

15 Tim Weiner, 'New model arm soldier rolls closer to battle', in *New York Times*, 16 February 2005, available at: http://www.nytimes.com/2005/02/16/technology/16robots.html (last visited January 2012).

16 Gary E. Marchant, Braden Allenby, Ronald Arkin, Edward T. Barrett, Jason Borenstein, Lyn M. Gaudet, Orde Kittrie, Patrick Lin, George R. Lucas, Richard O'Meara, Jared Silberman, 'International governance of autonomous military robots', in *The Columbia Science and Technology Law Review*, Vol. 12, 2011, pp. 272–315.

17 See K. Anderson and M. Waxman, above note 8.

18 Drew McDermott, 'Artificial Intelligence Meets Natural Stupidity', in J. Haugland (ed.), *Mind Design*, MIT Press, Cambridge, 1981, pp. 143–160.

understanding to the program. McDermott suggests, instead, using names such as 'G0034' and seeing if others are convinced that the program implements 'understanding'.

Ronald Arkin's work on developing a robot with an artificial conscience provides us with a strong case study to explore what happens when wishful mnemonics and a particular anthropomorphic perception of robots and emotion are applied. He states: 'I am convinced that they [autonomous battlefield robots] can perform more ethically than human soldiers are capable of.'[19] Notice that he does not say that humans could *use* robots in a more ethical manner. Instead, he directs us into the mythical trope that the robots themselves will perform more ethically. This can lead to the mistaken conclusion that robots are capable of moral reasoning in warfare in the same way as humans. Once this premise is in place, all manner of false inferences can follow that could impact on military planning for the future about how armed robots are deployed in civilian areas.

The same author states that:

> it is a thesis of my ongoing research for the U.S. Army that robots not only can be better than soldiers in conducting warfare in certain circumstances, but they also can be more humane in the battlefield than humans.[20]

But surely the suggestion that robots could be more humane on the battlefield than humans is an odd attribution to make about machines. Humans may apply technology humanely, but it makes no sense to talk of an inanimate object being *humane*. That is an exclusive property of being human. It implies that a robot can show kindness, mercy, or compassion or that it has humanistic values (robot compassion will be discussed in more detail below). The statement that robots can be more humane than humans leads to the very worrying implication that robots will humanize the battlefield when in fact they can only dehumanize it further.

This is not just being picky about semantics. Anthropomorphic terms like 'ethical' and 'humane', when applied to machines, lead us to making more and more false attribution about robots further down the line. They act as linguistic Trojan horses that smuggle in a rich interconnected web of human concepts that are not part of a computer system or how it operates. Once the reader has accepted a seemingly innocent Trojan term, such as using 'humane' to describe a robot, it opens the gates to other meanings associated with the natural language use of the term that may have little or no intrinsic validity to what the computer program actually does.

Several authors discussing robot ethics make a distinction between functional and operational morality.[21] Functional morality 'assumes that robots

19 See R. C. Arkin, above note 6, pp. 47–48.

20 Ronald C. Arkin, 'Ethical Robots in Warfare', in *IEEE Technology and Society Magazine*, Vol. 28, No. 1, Spring 2009, pp. 30–33.

21 E.g. Robin Murphy and David Woods, 'Beyond Asimov: the three laws of responsible robotics', in *IEEE Intelligent Systems*, Vol. 24, No. 4, July–August 2009, pp. 14–20; Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, New York, 2009.

have sufficient agency and cognition to make moral decisions'.[22] Operational morality is about the ethical use of robots by the people who make decisions about their use, who commission, handle, and deploy them in operational contexts.

In a recent report, the US Defense Advisory Board discusses the problems of functional morality citing Arkin's work and concludes by saying that:

> [t]reating unmanned systems as if they had sufficient independent agency to reason about morality distracts from designing appropriate rules of engagement and ensuring operational morality.[23]

To illustrate the distinction between robots being used ethically (operational morality) versus robots being ethical (functional morality), I will use the example of a thermostat. Consider an unscrupulous care home owner who saves money by turning down the heating in the winter, causing hypothermia in elderly residents. This is clearly unethical behaviour. As a result, the owner is legally forced to install a thermostat that is permanently set to regulate the temperature at a safe level. Would we want to describe the thermostat itself (or the heating system as a whole) as being ethical? If someone altered the setting, would we now say that it was behaving unethically?

The moral decision to have the thermostat installed was made by humans. This is operational morality. The thermostat is simply a device being used to ensure compliance with the regulations governing elder care. This is not so different from a robot in that both follow pre-prescribed instructions. Agreed, a robot is capable of some greater complexity, but it is inaccurate to imply that its programmed movements constitute ethical behaviour or functional morality. Yet when Arkin discusses emotion, it is in a way similar to the thermostat example here.

He states that, 'in order for an autonomous agent to be truly ethical, emotions may be required at some level'.[24] He suggests that if the robot 'behaves unethically', the system could alter its behaviour with an 'affective function' such as guilt, remorse, or grief.[25] Indeed, the way that he models guilt provides considerable insight into how his emotional terms operate as Trojan horses where the 'wished for' function of the label differs from the 'actual' software function.

He models guilt in a way that works similarly to our thermostat example. Guilt is represented by a 'single affective variable' designated V_{guilt}. This is just a single number that increases each time 'perceived ethical violations occur' (for which the machine relies on human input). When V_{guilt} reaches a threshold, the machine will no longer fire its weapon just as the thermostat cuts out the heat when the temperature reaches a certain value. Arkin presents this in the form of an

22 See R. Murphy and D. Woods, *ibid.*

23 Task Force Report, 'The Role of Autonomy in DoD Systems', Department of Defense – Defense Science Board, July 2012, p. 48, available at: http://www.fas.org/irp/agency/dod/dsb/autonomy.pdf (last visited January 2012).

24 See R. C. Arkin, above note 6, p. 174.

25 *Ibid.*, p. 91.

INTERNATIONAL REVIEW of the Red Cross

equation:

$$\text{IF } V_{guilt} > Max_{guilt} \text{ THEN } P_{l-ethical} = \text{ø}$$

where V_{guilt} represents the current scalar value of the affective state of Guilt, and Max_{guilt} is a threshold constant.[26]

This Trojan term 'guilt' carries with it all the concomitant Dostoevskian baggage that a more neutral term such as 'weapons disabler' would not. Surely, guilt minimally requires that one is aware of one's responsibilities and obligations and one is capable of bearing responsibility for one's actions. Of course the robot, with its thermostat-like guilt function, does not have this awareness, but this is exactly what the use of the word 'guilt' smuggles into the argument.

The Trojan term 'guilt' plays into the cultural myth of AI. Once this seemingly innocent 'affective' Trojan has been taken in, its doors open to beguile readers into accepting further discussions of the 'internal affective state of the system', 'affective restriction of lethal behaviour',[27] 'affective processing',[28] and how 'these emotions guide our intuitions in determining ethical judgements'.[29]

The same author then wishes us to accept that simply following a set of programmed rules to minimize collateral damage will make a robot itself compassionate:

> by requiring the autonomous system to abide strictly to [the laws of war] and [rules of engagement], we contend that it does exhibit compassion: for civilians, the wounded, civilian property, other non-combatants.[30]

This is like calling my refrigerator compassionate because it has never prevented my children from taking food or drinks when they needed them.

Given this collection of linguistic, emotional Trojan terms being applied to the functions of a computer program, it is hardly surprising that Arkin comes to the conclusion that robots could perform more ethically and humanely on the battlefield than humans. We must be wary of accepting such descriptive terms at face value and make sure that the underlying computational mechanisms actually support them other than in name only. To do otherwise could create a dangerous obfuscation of the technical limits of autonomous armed and lethal robots.

It is not difficult to imagine the impact on lawmakers, politicians, and the military hierarchy about the development and use of lethal autonomous robots if they are led to believe that these machines can have affective states, such as guilt and compassion, to inform their moral reasoning. The mythical theme of the 'ethical robot soldier' being more humane than humans has spread throughout the media and appears almost weekly in the press. These terms add credence to the notion that there is a technological fix around the corner that will solve the moral problems of

26 *Ibid.*, p. 176.
27 *Ibid.*, p. 172.
28 *Ibid.*, p. 259.
29 *Ibid.*, p. 174.
30 *Ibid.*, p. 178.

automating lethality in warfare. This stumbling block to prohibition presents a terrifying prospect.

One of Arkin's stated motivations for developing an 'ethical' robot, and it is well meaning, is a concern for the unethical behaviour of some soldiers in warfare. He provides several examples and was disconcerted by a report from the Surgeon General's Office on the battlefield ethics of US soldiers and marines deployed in Operation Iraqi Freedom.[31] However, even if warfighters do sometimes behave unethically, it does not follow that technological artefacts such as robots, that have no moral character, would perform more ethically outside of mythical AI. When things go wrong with humanity it is not always appropriate to just reach for technology to fix the problems.

The young men and women who fight our wars are capable of being ethical in their own lives. We must ensure that their moral reasoning capabilities are translated and used for the difficult situations they find themselves in during battle. Rather than funding technological 'hopeware', we need to direct funding into finding out where and when warfighters' ethical reasoning falls down and provide significantly better ethical training and better monitoring and make them more responsible and accountable for their actions. It is humans, not machines, who devised the laws of war and it is humans, not machines, who will understand them and the rationale for applying them.

Prohibiting the development of lethal autonomy

Legal advisors should not be distracted by the promise of systems that may never be possible to implement satisfactorily. It is vital that legal advice about autonomous armed robots is not polluted by anthropomorphic terminology that promises technological fixes. Advice about the indiscriminate nature of autonomous armed robots should come upstream and early enough to halt costly acquisition and development programs. As suggested by McClelland:

> it is important that the provision of formal written legal advice be synchronized with the acquisition process. If it is not, then there is a real danger that the legal advice will not be considered adequately in key decisions regarding the future acquisition of the equipment.[32]

Under IHL, there is no requirement for machines to be ethical or humane. The requirement is that they be used with appropriate restraint and respect for humanity.[33] In my view, given the severe limitations of the control that can be

31 *Ibid.*, p. 47.

32 Justin McClelland, 'The review of weapons in accordance with Article 36 of Additional Protocol', in *International Review of the Red Cross*, Vol. 85, No. 850, 2003, pp. 397–415.

33 I am uncomfortable with this expansion of the automation of killing for a number of other reasons that there is not space to cover in this critique. See, for example, Noel E. Sharkey, 'Saying — No! to Lethal Autonomous Targeting', in *Journal of Military Ethics*, Vol. 9, No. 4, 2010, pp. 299–313.

INTERNATIONAL REVIEW of the Red Cross

engineered into autonomous lethal targeting of humans, armed autonomous robots should be banned in the same way as other indiscriminate weapons.[34]

It could be argued that there are already weapons laws in place, such as Article 36 of Additional Protocol I.[35] But with the current drive towards autonomous operation, why has there not yet been any state determination as to whether autonomous robot employment, in some or all circumstances, is prohibited by Protocol I? This is a requirement of Article 36 for the study, development, acquisition, or adoption of any new weapon.[36] The 1980 Convention on Certain Conventional Weapons (CCW) also fits the bill. It bans weapons such as blinding laser weapons.[37] The aim is to prohibit weapons whose harmful effects could spread to an unforeseen degree or escape from the control of those who employ them, thus endangering the civilian population.

The reason why Article 36 may not have been applied and why autonomous lethal robots would be hard to get onto the CCW list is most likely because autonomous robots are not weapons systems until they are armed. Even locating people (targeting) does not make them weapons. It would only be possible to include them on the list after they have been developed which may then be too late. The worry is that arming an autonomous robot system will be a relatively simple add-on once the other technologies are in place. It is not difficult to repurpose a robot for combat as we have seen with the arming of the Predator drone in February 2001.[38]

Regardless of current intentions, if one state gains strong military advantage from using armed lethal autonomous robots, what will inhibit other states, in danger of losing a war, from following suit? We only have to look at the International Court of Justice decision, or more properly non-decision, on nuclear weapons[39] to realize how easy it would be to use autonomous lethal targeting, whether it was provably discriminate or not. The Court ruled that, in the current state of international law and given the facts at its disposal, it was not possible to conclude definitively whether the threat or use of nuclear weapons would be lawful or unlawful in extreme circumstances of self-defence (circumstances in which the very survival of the defending state would be at stake).[40] It would not be too fantastic to imagine the phrase 'autonomous armed robots' being substituted for 'nuclear weapons'. Armed robots seem a lesser beast than nuclear weapons unless they are

34 See also the statement of the International Committee for Robot Arms Control (ICRAC), at the Berlin Expert Workshop, September 2010, available at: http://icrac.net/statements/ (last visited 1 June 2012).

35 Additional Protocol I. This was not signed by the US.

36 There is a touch of the hat to the idea that there may be ethical issues in the 'unmanned systems integrated road map 2009–2034', but no detailed studies of the law or the issues are proposed.

37 United Nations Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects, in force since 2 December 1983 and an annex to the Geneva Conventions of 12 August 1949. I thank David Akerson for discussions on this issue.

38 Walter J. Boyne, 'How the predator grew teeth', in *Airforce Magazine*, Vol. 92, No 7, July 2009, available at: http://bit.ly/RT78dP (last visited January 2012).

39 International Court of Justice (ICJ), *Legality of the Threat or Use of Nuclear Weapons*, Advisory Opinion of 8 July 1996, available at: http://www.icj-cij.org/ (last visited 16 May 2012).

40 *Ibid.*, para. 105, subpara. 2E.

armed with nuclear weapons. So the substitution is easy. However, it is likely that it would take much less than the imminent collapse of a state before indiscriminate autonomous robots were released. Without an explicit ban, there is an ever-increasing danger that military necessity will dictate that they are used, ready or not.[41]

Nation states are not even discussing the current robot arms race. The only international instrument that discusses unmanned armed vehicles (UAVs) is the Missile Technology Control Regime (MTCR) established in 1987. This is a network of thirty-four countries that share the goal of preventing the proliferation of unmanned delivery systems for weapons of mass destruction. It is more concerned with missiles, but restricts export of UAVs capable of carrying a payload of 500 kilos for at least 300 kilometres. It is not overly restrictive for armed drones such as the Predator and does little to prevent their proliferation.

The MTCR is voluntary and informal with no legal status. It has been suggested that if the MTCR changed from a voluntary regime to a binding regime, further proliferation could be addressed by international law.[42] However, the MTCR currently only restricts export of a certain class of armed drones and does nothing to restrict their deployment. Moreover, US military contractors have lobbied to have export restrictions loosened to open foreign markets. On 5 September 2012, the Department of Defense announced new guidelines to allow sixty-six unspecified countries to buy American-made unmanned air systems.

Perhaps the most promising approach would be to adopt the model created by coalitions of non-governmental organizations (NGOs) to prohibit the use of other indiscriminate weapons. The 1997 mine-ban treaty was signed by 133 nations to prohibit the use of anti-personnel mines and 107 nations adopted the 2008 Convention on Cluster Munitions in 2008. Although a number of countries including the US, Russia, and China did not sign these treaties, there has been little substantial use of these weapons since and the treaty provisions could eventually become customary law.

Conclusion

It is incumbent upon scientists and engineers in the military context to work hard to resist the pressure of the cultural myth of robotics and to ensure that the terminology they use to describe their machines and programmes to funders, policymakers, and the media remains objective and does not mire them and others in the mythical. They must be wary of descriptive terms that presuppose the functionality of their programs (e.g. ethical governor, guilt functions, etc.) and consider the impact that such descriptions will have on the less technical.

41 See N. Sharkey, above note 2.

42 Valery Insinna, 'Drone strikes in Yemen should be more controlled, professor says', interview with Christopher Swift for the *National Defence Magazine*, 10 October 2006, available at: http://tinyurl.com/8gnmf7q (last visited January 2012).

Such terms can create unfounded causal attributions and may confuse proper discussion of the IHL issues.

It is important that the international community acts now while there is still a window of opportunity to stop or, at the very least, discuss the control and limits of the robotization of the battlespace and the increasing automation of killing. In my opinion, a total global ban on the development of autonomous lethal targeting is the best moral course of action. I have argued here that notions about ethical robot soldiers are still in the realms of conjecture and should not be considered as a viable possibility within the framework necessary to control the development and proliferation of autonomous armed robots. Rather than making war more humane and ethical, autonomous armed robotic machines are simply a step too far in the dehumanization of warfare. We must continue to ensure that humans make the moral decisions and maintain direct control of lethal force.