Wishful Mnemonics and Autonomous Killing Machines

Since 19th century military theorist Carl von Clausewitz first coined the phrase 'the fog of war'[1], the problem of how adequately to interpret unfolding events in the field of battle has been placed explicitly at the centre of military affairs. At the same time, 20th century developments in military technologies towards increasingly 'network-centric' warfare, along with accelerating initiatives in battlefield automation, have resulted in ever more tightly coupled systems of situation assessment and response[2]. While the premise that the deployment of information and communications technologies would help to dispel the uncertainties of warfare is now in question[3], developments in battlefield automation have continued in the direction of increasingly autonomous systems.

Autonomy and accountability in warfare

Under current military policy, the deployment of armed robots requires that human operators take decisions on the application of lethal force. Over the last decade, however, the 'Roadmaps' of United States forces have made clear the desire and intention to develop and use autonomous battlefield robots[4]. The US Department of Defense *Unmanned Systems Integrated Roadmap 2011–2036* describes the advantages of autonomous over existing automatic systems:

"Dramatic progress in supporting technologies suggests that unprecedented levels of autonomy can be introduced into current and future unmanned systems... Automatic systems are fully preprogrammed and act repeatedly and independently of external influence or control... However, the automatic system is not able to define the path according to some given goal or to choose the goal dictating its path. By contrast, autonomous systems are self-directed toward a goal in that they do not require outside control, but rather are governed by laws and strategies that direct their behavior... The special feature of an autonomous system is its ability to be goal-directed in unpredictable situations. This ability is a significant improvement in capability compared to the capabilities of automatic systems"[5].

While there are assurances that "[f]or the foreseeable future, decisions over the use of force and the choice of which individual targets to engage with lethal force will be retained under human control in unmanned systems"[6], these are countered by the emphasis throughout these reports on the benefits of increased autonomy, and research and development aimed at taking the human out of the control loop is well underway. The end goal is a network of aerial, land, and underwater robots that will operate together autonomously to locate their targets and destroy them without human intervention. The US is not the only country, moreover, with autonomous robots in

their sights: China, Russia, Israel and the UK are following suit.

At the same time, nation states engaged in armed conflict remain accountable in principle to the requirements of International Humanitarian Law (IHL)[7]. A major question that arises within this legal framework is the ability of autonomous armed robot systems to distinguish between combatants and non-combatants, or other protected actors such as combatants who are wounded or have surrendered. There are systems currently in use that have a weak form of discrimination. The Israeli Harpy, as one example, is a loitering munition that detects radar signals. When it finds one, it references its database to determine if the signal is friendly and if not, it targets the radar. This type of discrimination relies, however, on the accuracy of the database, and fails as well to take into account the context of the signal; for example, whether the radar is positioned on an anti-aircraft station, or on the roof of a school or a hospital[8].

In the current state of the art, robots lack three components required to ensure compliance with International Humanitarian Law. The first concerns the Principle of Distinction[9], which would require that robots have adequate vision or other sensory processing systems for separating combatants from civilians, particularly in circumstances where the former are not in uniform, and for reliably differentiating wounded or surrendering combatants from those who pose an imminent threat. Sensors such as cameras, infrared, sonars, lasers, temperature sensors, ladars and the like may be able to tell us that something is a human, but they cannot tell us much else. There are systems currently in the labs that can recognize still faces matched against a database, and they might eventually be deployed for individual targeting in limited circumstance. But British teenagers beat surveillance cameras simply by wearing hooded jackets. And how accurate will facial recognition systems be with moving targets, or targets tracked dynamically from the air?

The more basic problem in meeting the requirements of the Principle of Distinction is that we do not have an adequate definition of a civilian that can be translated into a recognition algorithm. Nor can we get one from the Laws of War. The 1949 Geneva Convention requires the use of 'common sense,' while the 1977 Protocol I essentially defines a civilian in the negative sense, as someone who is not a combatant[10]. Even if machines had adequate sensing mechanisms to detect the difference between civilians and uniform-wearing military, they would fail under situations of contemporary warfare where combatants are frequently not in uniform. While robotics may move towards some limited sensory and visual discrimination in certain narrowly constrained circumstances within the next 50 years, human level discrimination with adequate common sense reasoning and situational awareness may prove computationally intractable[11]. At this point, at least, there is no evidence or research results to suggest otherwise.

A second IHL issue is the Principle of Proportionality[12]. One robotics

expert has argued that robots could calculate proportionality better than humans[13]; however this concerns what we might call the easy proportionality problem: that is, minimising collateral damage by choosing the most appropriate weapon or munition and directing it accurately according to a specified target. 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 strike benefit military objectives, or hinder them because of its effects on the local civilian population? The list of questions is open-ended. It is a qualitative judgment regarding what cost in civilian injury is proportional to direct military advantage. It is imperative that such decisions are made by responsible, accountable human commanders who can weigh the options based on experience and on adequate situational awareness. 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'[14].

A third issue, which cuts across these two, is that of accountability. A robot does not have moral agency and consequently cannot be held accountable for its actions. Robert Sparrow[15] argues that irresolvable ambiguities surrounding questions of responsibility for actions taken in the case of artificially intelligent robotic weapons (particularly in relation to the automation of target identification) render their deployment irremediably unethical. Anderson and Waxman[16] dismiss the accountability objection out of hand, on the grounds that 'post-hoc judicial accountability in war is just one of many mechanisms for promoting and enforcing compliance with the laws of war'. But at the least the question of responsibility is vastly complicated in the case of autonomous robot weapons, and deploying a weapon without a clear chain of accountability is not a morally defensible option.

It is on the basis of these three concerns that we call for a ban on autonomous lethal targeting by robots[17]. A major stumbling block to a prohibition on the development of armed autonomous robots, however, is the claim by proponents of lethal autonomous robots that there are technological 'fixes' that will make them behave more ethically and more humanely than soldiers on the battlefield. We argue that this has more to do with the language being used to describe robots, than with what robots can actually do.

Anthropomorphism and wishful mnemonics in AI

Robots have been depicted in science fiction, in media reporting, and by some robotics experts as sentient machines that can reason and act in ways superior to humans, as well as 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[18]. We are all susceptible to this. Journalists are particularly caught up in these forms of attribution, as they know that their readers love it. Within the field of AI and robotics as well, it is acceptable and even customary to describe robots with an anthropomorphic narrative. While this can be harmless in casual conversations in the lab, it is a perilous basis for legal and political discussion about enabling autonomous lethal machines.

In an influential paper, Drew McDermott, 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[19]. Such terms represent a researcher's aspirations, he argues, 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', by misattributing understanding to the program. McDermott suggests, instead, that we use names such as 'G0034,' and then see if it is as easy to argue that the program implements 'understanding'.

The combination of anthropomorphism and wishful mnemonics, we would suggest, underwrites the programme of roboticist Ronald Arkin, who states: '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]. Anthropomorphic terms like 'humane', when applied to machines, carry along with them a rich, interconnected web of concepts that are not technically part of a computer system or how it operates. We need to ask: How would 'humaneness' be specified programmatically, and then matched appropriately to an open horizon of contingent situations?[21]

While Arkin cites lack of fear as one element that could ensure the greater humanity of battlefield robots, he also states that 'in order for an autonomous agent to be truly ethical, emotions may be required at some level' [22]. More specifically, he suggests that if the robot 'behaves unethically', the system might alter its behaviour with an 'affective function' such as guilt, remorse or grief[23]. Arkin models guilt as a 'single affective variable' designated $Vguilt$. This is a single number that increases each time 'perceived ethical violations occur' (for which the machine relies on human input). When $Vguilt$ reaches a threshold, the machine will no longer fire its weapon, just as a thermostat cuts out the heat when the temperature reaches a certain value. Arkin presents this in the form of an equation:

```
IF Vguilt > Maxguilt
THEN Pl-ethical = 0
```

where $Vguilt$ represents the current scalar value of the affective state of guilt, and $Maxguilt$ is a threshold constant[24]. This term, guilt, carries with it all of the connotations that a more neutral term, such as 'weapons disabler', would not.

Arkin assumes, inter alia, that the Laws of Armed Conflict and Rules of Engagement resolve questions of ethical conduct in war fighting, and could be effectively encoded within the control architecture of a robotic system[25]. Arkin then wishes us to accept that 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 noncombatants'[26]. Peter Asaro, in contrast, in considering the programmability of the laws of war, draws on Just War Theory, the principles underlying most of the international laws regulating warfare, including the Geneva and Hague Conventions[27]. Asaro reminds us that the Laws of Armed Conflict comprise what he characterizes as a 'menagerie' of international laws and agreements (such as the Geneva Conventions), treaties (such as the anti-personnel landmine ban), and domestic laws, and the Rules of Engagement (ROE) rest on the principles of discrimination and proportionality. As Asaro explains; 'the ROE are devised to instruct soldiers in specific situations, and take into account not only legal restrictions but also political, public relations, and strategic military concerns... They often appear ambiguous or vague to the soldiers on the ground who observe situations that do not always fall neatly into the distinctions made by lawyers', while the Principle of Proportionality is 'abstract, not easily quantified, and highly relative to specific contexts and subjective estimates of value'[28]. These are far from algorithmic specifications for decision-making and action, in other words, not least (as in the case of recent contests over who is protected under the Geneva Conventions) over the identification of a 'combatant.'

We must be wary, in sum, of accepting 'wishful mnemonics' at face value, ensuring rather that the underlying computational mechanisms actually support the functions named, in other than name only. To do otherwise could result in a dangerous obfuscation of the actual technical limits of autonomous armed and lethal robots. It is not difficult to imagine the impact on lawmakers, politicians and military decision-makers if they are led to believe that lethal autonomous robots can have affective states such as guilt and compassion to inform their moral reasoning. The premise of the 'ethical robot soldier' being more humane than humans has spread throughout the media and appears almost weekly in the press. These representations add credence to the notion that there is a technological fix around the corner that will solve the real moral problems of unethical behaviour in warfare, through the automation of lethality. Rather than hoping for technological solutions, we need to direct attention and funding to understanding under what conditions the legal and ethical reasoning of human soldiers fails in warfare, and work to mitigate those conditions as well as to provide better training, closer monitoring and greater responsibility and accountability for military actions.

Prohibiting the development of lethal autonomy

It is our position that discussion about the limitations and risks of autonomous armed robots should come upstream and early enough to halt costly acquisition and development programs. It could be argued that there are already relevant weapons laws in place, such as Article 36 of Additional Protocol I[29]. 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. Bolton, Nash and Moyes[30] argue for the relevance of this legal framework to a ban on autonomous armed robots, in terms of their comparability to anti-personnel landmines with respect to problems of autonomy and inadequate discrimination of their targets:

"In banning anti-personnel landmines the global humanitarian community acted to address a military technology that has caused extensive suffering to civilians, but is also a weapon type that raises particular moral concerns because of the way in which it functions... Weapons that are triggered automatically by the presence or proximity of their victim can rarely be used in a way that ensures distinction between military and civilian".

These questions are made more urgent insofar as, if one state gains strong military advantage from using armed lethal autonomous robots, there is little to inhibit other states from following suit. Yet nation states are not even discussing the current robot arms race. On the contrary, US military contractors have lobbied to have export restrictions loosened to open foreign markets. On September 5th, 2012, the Department of Defense announced new guidelines to allow 66 unspecified countries to buy American-made unmanned air systems.

Perhaps the most promising approach would be to adopt the model created by coalitions of 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 Convention on Cluster Munitions in 2008. Although a number of countries including the U.S., 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, particularly in the military context, to work to ensure that the terminology that they use to describe their systems to funders, policy makers and the media does not resort to unsubstantiated anthropomorphism or wishful mnemonics. We must be wary of evocative terms that imply the functionality of programs (e.g., ethical governor, guilt functions, etc.) rather than provide technical descriptions of actually-existing capabilities. More generally, 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 robotisation of the battlespace and the increasing automation of killing. In our view a global ban on the development and deployment of autonomous lethal targeting is the best course of action, both legally and morally. We have argued here that notions about ethical robot soldiers are still in the realm 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 systems comprise a step too far in the automation, and associated dehumanization, of warfare. Rather than turning to further automation in the face of the intensifying uncertainties of warfare, and the persistent occurrence of extra- or illegal actions in the conduct of killing, we must renew our efforts to ensure that humans are held responsible for decisions regarding the use of violent force upon other human beings.

Footnotes

[1] Clausewitz, Carl von. (1976) *On War*. Michael Howard and Peter Paret (trans, and ed.) Princeton, NJ: Princeton University Press.

[2] On the history of these developments see Paul Edwards (1997) *The Closed World*, Cambridge, MA: MIT Press, and Agatha Hughes and Thomas Hughes (2011) *Systems, Experts and Computers*. Cambridge, MA: MIT Press. On network-centric warfare see James der Derian (2009) *Virtuous War*, New York: Routledge.

[3] See for example Patrick Cronin (2008) *The impenetrable fog of war: reflections on modern warfare and strategic surprise*. Westport, CT: Praeger Security International.

[4] See US Department of Defense, *Unmanned Systems Integrated Roadmap FY2011-2036, Reference Number 11-S-3613*, 2011; United States Air Force *Unmanned Aircraft Systems Flight Plan 2009-2047*, Headquarters of the United States Air Force, Washington, DC, 18 May 2009; Ministry of Defence The UK *Approach to Unmanned Aircraft Systems, Joint Doctrine Note 2/11*, 30 March, 2011.

[5] US DOD *Unmanned Systems Integrated Roadmap FY2011-2036*, p. 43.

[6] US DOD *Unmanned Systems Integrated Roadmap FY2011-2036*, p. 17. Department of Defense *Directive Number 3000.09*, November 21, 2012 offers the DoD's most recent qualifications on autonomy, but for a response see Noel Sharkey, *America's mindless killer robots must be stopped* Guardian Commentary
http://www.guardian.co.uk/commentisfree/2012/dec/03/mindless-killer-robots
3 December 2012 (accessed 19 January 2013). See also Noel Sharkey, *Cassandra or the false prophet of doom: AI robots and war*, IEEE Intelligent Systems, Vol. 23, No. 4, 2008, pp. 14–17.

[7] Also known as the Law of War or the Law of Armed Conflict, IHL "is a set of rules which seek, for humanitarian reasons, to limit the effects of armed conflict. It protects persons who are not or are no longer participating in the hostilities and restricts the means and methods of warfare". http://www.icrc.org/eng/resources/documents/legal-fact-sheet/humanitarian-law-factsheet.htm (accessed 19 January 2013).

[8] Other systems currently in use can be seen as precursors to autonomy; for a relevant list see Human Rights Watch Losing Humanity: The case against killer robots, 2012,

http://www.hrw.org/news/2012/11/19/ban-killer-robots-it-s-too-late.

[9] For a definition see: (accessed 19 January 2013)
http://www.icrc.org/customary-ihl/eng/docs/v1_cha_chapter1_rule1.

[10] Art 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).

[11] See Noel E. Sharkey, "Grounds for Discrimination: Autonomous Robot Weapons", in RUSI Defence Systems, Vol. 11, No. 2, 2008, pp. 86-89. Situational awareness is defined as "understanding of the operational environment in all of its dimensions– political, cultural, economic, demographic, as well as military factors." Dostal, Major Brad C. (2001). *Enhancing situational understanding through the employment of unmanned aerial vehicles.* Center for Army Lessons Learned. Retrieved from (accessed 19 January 2013)
http://www.globalsecurity.org/military/library/report/call/call_01-18_ch6.htm.

[12] See (accessed 19 January 2013)
http://www.icrc.org/customary-ihl/eng/docs/v1_cha_chapter4_rule14.

[13] Ronald C. Arkin (2009) *Governing Lethal Behavior in Autonomous Systems*, Taylor-Francis, pp. 66–68.

[14] Mark Anderson (2010) *How Does a Terminator Know When to Not Terminate*, Discover Magazine, p. 40. While it is clear that humans themselves frequently fail in these assessments, the logical corrollary of this is not, in our view, a justification for further automation of decision making. We return to this issue below.

[15] Robert Sparrow (2007) *Killer robots.* Journal of Applied Philosophy, 24, pp. 62Ð77. See also Armin Krishnan (2009) *Killer Robots: Legality and Ethicality of Autonomous Weapons*, Surrey, UK: Ashgate Publishing Limited, and Jutta Weber (2009) *Robotic Warfare: Human Rights and the Ethics of Robotic Machines* in R. Capurro and M. Nagenborg (eds.) Ethics and Robotics. Heidelberg: AKA Verlag, pp. 83–103.

[16] Kenneth Anderson and Matthew Waxman 2012 *Law and Ethics of Robot Soldiers*, Policy Review.

[17] 'The Scientists' Call to Ban Autonomous Lethal Robots', available for signing at http://icrac.net/call/.

[18] See Amanda Sharkey and Noel Sharkey (2006) *Artificial Intelligence and Natural Magic*, Artificial Intelligence Review, Vol. 25, No 1–2, pp. 9–19; Jackie Stacey and Lucy Suchman *Animation and Automation: The liveliness and labours of bodies and machines.* Body & Society 18 (1): 1–46.

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

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

[21] See Lucy Suchman (2007) *Human-Machine Reconfigurations: Plans and situated actions, expanded edition*, Cambridge University Press.

[22] See above note 13, p. 174.

[23] Ibid., p. 91.

[24] Ibid., p. 176.

[25] On the Laws of Armed Conflict see note 9 above. The Rules of Engagement are "Directives issued by competent military authority that delineate the circumstances and limitations under which United States forces will initiate and/or continue combat engagement with other forces encountered" http://www.cc.gatech.edu/ tpilsch/AirOps/cas-roe.html

(accessed 19 January 2013).

[26] See above note 13, p. 178.

[27] Peter Asaro (2009) *How just could a robot war be?* In P. Brey, A. Briggle, & K. Waelbers (Eds.), Current Issues in Computing And Philosophy (pp. 50Ð64). Amsterdam: IOS Press, pp. 50–64.

[28] Peter Asaro (2009) *Modeling the Moral User.* IEEE Technology and Society Magazine, 2009, p. 21.

[29] See footnote 10 above regarding Additional Protocol I, though note that this has not been signed by the U.S.

[30] Matthew Bolton, Thomas Nash and Richard Moyes (2013) *Ban Autonomous Armed Robots* http://www.article36.org/statements/ban-autonomous-armed-robots/, (accessed 19 January 2013).

Noel 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.

Lucy Suchman is Professor of Anthropology of Science and Technology at Lancaster University in the UK.