The Bright Red Line of Responsibility

Mark R. Waser
Digital Wisdom Institute
MWaser@DigitalWisdomInstitute.org

Abstract

The last six months has seen a rising tsunami of interest in "killer robots" and "autonomy" in weapons systems. We argue that most of the "debate" has been derailed by emotionally inflammatory terms, red herrings facilitated by the overuse of "suitcase" words rather than precisely defined terms and entirely spurious arguments. There have been numerous proposals for "robot arms control" but we contend that such framing is a major part of the problem even when done by responsible parties. Our proposal is to move forward by referring back to the one necessary but simple core concept of responsibility.

1 Introduction

On November 19, 2012, Human Rights Watch (HRW) issued a 50-page report "Losing Humanity: The Case against Killer Robots" (Human Rights Watch, 2012) outlining concerns about "fully autonomous weapons that could select and engage targets without human intervention" and claiming that a "preemptive prohibition on their development and use is needed". Two days later, the United States Department of Defense released Directive 3000.09 (US DoD, 2012) "for the development and use of autonomous and semi-autonomous functions in weapon systems". Now, the International Committee for Robot Arms Control (ICRAC) has issued a Scientists' Call (ICRAC, 2013) to Ban Autonomous Lethal Robots "in which the decision to apply violent force is made autonomously".

We believe that this call is dangerously naïve because its specific goal distracts from its stated and true goal. It is far too tempting to argue over possibilities, potentialities and existential fears – rather than buckling down and hammering out the past due assignment of human responsibility and accountability. There have been numerous proposals for "robot arms control" (Asaro, 2008; Altmann, 2009; Krishnan, 2009; Sparrow, 2009; Marchant, et al., 2011; Sparrow, 2011; Wallach & Allen, 2012) but we strongly proclaim that this is far too broad an area to make any overdue progress in a short period of time. The *only* way to dig ourselves out of the hole of already deployed clear and present dangers is to address these issues through the focusing lens of human responsibility and accountability. Indeed, we would even go so far as to argue that the subject of morality itself is an overly-attractive red herring to short-term progress. Correctly settle the issues of responsibility and accountability and the general public will handle what is acceptable and what is not acceptable.

2 Undermining Responsibility

The ICRAC starts its justification for a ban by stating "*We are concerned about the potential of robots to undermine human responsibility in decisions to use force, and to obscure accountability for the consequences.*" Yet, they themselves are already clouding the issue by phrasing it as the "potential of robots" instead of "humans using robots as an excuse" to undermine and obscure.

Instead of clearly asking and answering the question "Where does human responsibility and accountability end?" they repeatedly undermine and obscure their own stated purpose with entirely spurious arguments about robot weapons' capabilities followed by the "question whether they ***could*** [emphasis ours] meet the strict legal requirements for the use of force" as if the machines themselves were responsible for meeting the requirements.

Much of the problem is that there are innumerable examples which do "meet the strict legal requirements" and an equally large (or larger) number of examples which do not. Critically, however, the difference is virtually always based upon the circumstances under which they are *used* – generally because the robot does not have the proper sensors and algorithms for a specific environment (Guarini & Bello, 2011) and cannot be overridden by a human when necessary – rather than never having a legal use (an instance where a ban would become appropriate). Since robots are still tools rather than entities (although we will address that transition shortly), the responsible human being using that tool or the party responsible for allowing the tool to pass out of responsible hands are the ones who should be accountable for that use. If a human being initiates use of a robot under illegal conditions or with the reasonable possibility of transition into illegal conditions – the human being initiating that use (or the entities who allowed him to come into possession and control of the robot) are entirely responsible for the consequences of that initiation.

Unfortunately, ICRAC personnel insist (Sharkey, 2011) on repeatedly obscuring these simple facts of responsibility with spurious arguments. After conceding that "According to the laws of war, a robot could potentially be allowed to make lethal errors, providing that noncombatant casualties were proportional to the military advantage gained", the immediately question is "how is a robot supposed to calculate . . . ." The problem is that any details of the calculation are entirely irrelevant except for how their success percentages and failure modes affect the behavior of the responsible party placing the robot in a given situation. Yet, once again, the focus is improperly placed upon the robot.

Next, it is claimed that "it would be difficult to allocate responsibility in the chain of command or to manufacturers, programmers, or designers – and being able to allocate responsibility is essential to the laws of war". This is entirely incorrect. Responsibility *must* fall upon the chain of command. It is entirely unacceptable that it be possible for the responsibility for starting a war to fall upon a manufacturer. The manufacturer is certainly liable under the civil laws governing product liability (Asaro, 2011) and the chain of command can easily extract its pound of flesh for any "failure to warn" or "failure to take proper care", but it is *entirely* contingent upon the chain of command to take responsibility for the use of any product (Champagne & Tonkens, 2012) and that initial authorization responsibility begins at the top seems perfectly clear to the US military (US Air Force, 2009):

> *Authorizing a machine to make lethal combat decisions is contingent upon political and military leaders resolving legal and ethical questions. These include the appropriateness of machines having this ability, under what circumstances it should be employed, where responsibility for mistakes lies and what limitations should be placed upon the autonomy of such systems.*

3 A Clear and Present Danger

So let us move on to the specific example of the long-deployed Israel Aerospace Industries Harpy, a fire-and-forget "loitering attack" UAV, designed to autonomously fly to and patrol an assigned area and attack any hostile radar signatures with a high explosive warhead. Noel Sharkey claimed in the HRW report that "it cannot distinguish between an anti-aircraft defense system and a radar placed on a school by an enemy force". This is particularly problematical since the Harpy is meant/designed to be pre-emptive and suppressive in that you can assign it to a patrol area with no radar and it will self-assign *ANY* target that appears without *any* opportunity for human intervention.

Who is going to take the blame if some "terrorist" (or injured Palestinian) does "spoof" an Israeli Harpy into obliterating an Israeli school? We would argue that it is the citizens of Israel who would be responsible with the highest levels of the Israeli government being the accountable parties. As unfair as it seems to "blame the victim" (and we don't deny the responsibility of the "spoofer"), this is a situation created entirely by the Israeli government. It actively pushed for and participated in the creation and deployment of the weapon and did not even ensure that it was safe for their own citizens when used under their own rules of engagement.

A much nastier case arises when someone reportedly steals a Harpy from China and it is used against a busy Japanese commercial airport. The Chinese government is clearly the accountable party but it is unclear what repercussions they should suffer. There is also the (correct) temptation to blame the Israeli government as well for allowing the sale of the Harpy to China (an action which the US has already "punished" by temporarily removing and then reinstating its status as Security Cooperative Participant in the Joint Strike Fighter program).

We would argue as well that the existence of the Harpy underscore the dangerous naïveté of those who do not share the "presumption that lethal autonomy is inevitable" unless that is merely part of an attempt at delaying the inevitable. Thirty-plus years of attempting to forge arms control treaties for cruise missiles (Gormley, 2008) should give a good indication of the likelihood of the genie being put back into the bottle. On the other hand, militaries absolutely hate not being in control and the second version of the Harpy, the larger IAI Harop, can be controlled in flight by a remote operator who can select targets via its electro-optical sensor – but this is in addition to, rather than in place of, the same anti-radar homing system.

4 "Suitcase" Words

Marvin Minsky has provided (Minsky, 2006) the useful concept of "suitcase" words – packed with many different meanings potentially only having the common feature "that each of them serves some goals of a person who packed them into that bag!" Such words "incorporate ambiguities that have evolved over centuries, to serve many important purposes – but also, they often handicap us by preserving outdated concepts." Autonomy and morality are just the most obviously problematical of the numerous examples that plague these debates – because they allow the improper conflation of present-day and far future circumstances that literally cannot exist simultaneously.

4.1 Autonomy

There is a vast spectrum between present-day operational autonomy which includes the ability to "select and engage targets without human intervention" and distant future "intentional autonomy" where the weapon/robot itself is truly making a *decision* to apply violent force. Indeed, we have argued previously (Waser, Dissecting the Scientists' Call to Ban Autonomous Lethal Robots, 2013) that HRW and ICRAC "*totally (and irresponsibly) conflate the entirely transparent, deterministic and reproducible following of clearly and rigidly defined algorithms of current systems with the non-repeating decision-making of some future self-modifying self-willed entity*" by "using misleading terms like "decisions" and "delegated" that imply some sort of volition or, worse, the possibility that responsibility *could* be delegated – presumably to take advantage of the fear of "terminator"-style "killer robots" (as in the subtitle of HRW's report)". ICRAC's Noel Sharkey has responded (Sharkey, Dissecting the Scientists' Call to Ban Autonomous Lethal Robots - Comments, 2013) that "Any wording attributing 'will' or high-level cognition to machines in the statement is unintentional" and that "There is no notion of self-willed or any kind of will involved except for the militaries who will use them." Yet, discussions about the ban invoke self-willed machines far more often than not.

Much of this is undoubtedly due to the fact that the HRW report repeatedly instigates a fear of self-willed machines – starting with its pejorative subtitle regarding “killer robots”. It has been argued (Lokhorst & van den Hoven, 2011) that if robot soldiers under lawful authority deserve that negative appellation then by that same token, human soldiers might then be called "killers," or even "murderers." The HRW also repeatedly refers to concerns about the present day lack of emotion while ignoring the fact that many researchers are finding that the functional equivalent of emotions are likely necessary long before the future possibility of full autonomy.

4.2 Morality

Unfortunately, virtually everyone is kicking up the cloud of dust that surrounds the suitcase word morality and what is and is not morally “acceptable”. The ICRAC ban starts by claiming that “wide adherence to the prohibitions on biological and chemical weapons as well as anti-personnel land mines” proves “that not all weapons are acceptable”; and “that fully autonomous robots that can trigger or direct weapons fire without a human effectively in the decision loop are similarly unacceptable.” We would argue that the previous bans are solutions to “Tragedy of the Commons” situations where all parties were better off with the ban rather than the elimination of something that was somehow “unacceptable” (since a country’s military is arguably delinquent if it does not produce such problematical systems as long as other countries are likely to be doing so). This argument, of course, would be inconvenient for the ICRAC in those situations where the use of such robots might easily lead to fewer civilian casualties and it would be obvious that it is the ban that should be “unacceptable”.

Even more prone to inspiring distracting debate is the proposal (Arkin, 2009) that the laws of war and rules of engagement can be computerized in a manner that will make it possible for robotic soldiers to behave more ethically than their human counterparts. Particularly interesting is the contention (Beavers, 2011) that “if it is within our power to build a machine that can make human beings more moral, both individually and collectively, then we have a prima facie moral obligation to build it”. We agree that the subject is endlessly fascinating and critically important – but, until exploration does (or does not) bear fruit, it is merely a distraction from immediate concerns. Further, whenever discussing machine ethics, it is critically necessary to differentiate between current ethical-impact agents, currently limited but improving implicit ethical agents, future explicit ethical agents, and fully-autonomous full ethical agents (Moor, 2006) and to realize that “the near future of moral machines is not and cannot be the attempt to recreate full moral agency” (Allen & Wallach, 2011).

5 Self-Willed Robots

Of course, we would be delinquent if we did not address the movie-bred fear of human-hating “terminators”. First off, it just simply is not going to happen that a machine will “just wake up” (Cameron, 1984). While we are among the most optimistic in terms of the time-frame for sapient robots, it is very clear that “self-creating” (autopoietic) machines have numerous requirements and precursors that we are nowhere close to implementing. Indeed, there is simply no way currently in which someone could deliberately create such an entity without an effort on the scale of the Manhattan project.

It must also be pointed out that the creation of self-willed robots is entirely a matter of human responsibility and accountability. Thus, we are far more concerned about the possibility of such stirring of human fears and prejudices on delaying artificial intelligence and robotics development and the increased risks and lack of advantages thereby incurred than any likely fully-autonomous machine. Further, we are also far more concerned about fear and prejudice’s likelihood to lead to human refusal

to grant ethical standing to self-willed robots (Waser, 2012) than we are about any other possible source of robot malice (or even indifference) towards humanity.

The most interesting aspect of a self-willed robot is if we could design it with a clear top-most goal to be moral instead of the confused conflation of evolved drives that humans are still trying to resolve into some semblance of coherence and integrity. Such a robot would answer the paradox of MorMach, an all knowing moral machine, the ultimate oracle in all matters concerning ethics (Beavers, 2011) merely by being capable of error (and thus moral transgression) while, by design, clearly being incapable of immorality. Further, since it would be deterministic, it could give us more traction on the issues of "free will" to settle many of the current neurologically inspired challenges to the criminal justice system (Cashmore, 2010).

6 Making War Easier

A final concern expressed by HRW is that "the gradual replacement of humans with fully autonomous weapons could make decisions to go to war easier". They invoke experts to claim that "drones have already lowered the threshold for war, making it easier for political leaders to choose to use force" (Singer, 2009), that "wars without military fatalities would remove one of the greatest deterrents to combat" (Borenstein, 2008), and that "unmanned systems create both physical and emotional distance from the battlefield, which a number of scholars argue makes killing easier." ICRAC's Noel Sharkey also offers similar arguments about drones (Sharkey, 2011) as well as numerous arguments about how allowing distance (physical and conceptual) from combat makes killing easier (Sharkey, 2011).

An extreme fable showing the disadvantages of this type of distancing (Coon & Hamner, 1967) depicts two planets conducting an entirely-simulated war with the denizens of each being subject to execution via "disintegration booth" based upon the results of simulated attacks. Leaders are proud because they are avoiding the destruction and devastation of war – which has therefore continued for over 500 years. A better way of solving the problem is revealed when a third party is unwillingly drawn into the conflict and destroys one planet's simulators. The parable ends as planetary leaders, terrified at the probability of "real" war, call for a ceasefire and begin peace negotiations.

The lesson that we take from all this and Sharkey's statement that "military robots are the fruit of a long chain of weapons development designed to separate fighters from their foes", is not that we should ban military robots – for, undoubtedly, they would merely be replaced by some other development to make fighting easy – but that we need to make war more difficult and painful for our leaders. While we could easily see leaders proudly proclaiming their skill at protecting "their" population from destruction and devastation in order to stay in power (simply by watching the news), we would argue that the parable is unrealistic in that civilians would not long stand for a situation involving even a painless death when alternatives were available – and it is entirely unclear how peace could be portrayed as entirely unavailable unless the populace can be distracted by some other threat (which is what current leaders seem to excel at). Further, we would be seriously remiss if we did not point out that distance and "video game" graphics do not seem to aid many drone pilots in avoiding the consequences of combat (Saletan, 2008).

7 Summary

We are long past the point where discussing a ban on autonomous lethal weapons makes any sense – particularly when considering the long-deployed Harpy, more than three decades of lack of success in forge arms control treaties for the much narrower category of cruise missiles, and the fact that most

of the objections by HRW and the ICRAC were answered long before they were raised (Lin, Bekey, & Abney, 2008; Lin, Bekey, & Abney, 2009; Marchant, et al., 2011). We need to stop debating the fascinating possibilities, potentialities and existential fears and start buckling down and hammering out the past due assignment of human responsibility and accountability. Investigation into the issues of autonomy and morality are critically important as well but they cannot be allowed to derail progress on issues that should have been resolved long ago.

References

Allen, C., & Wallach, W. (2011). Moral Machines: Contradiction in Terms or Abdication of Human Responsibility? In P. Lin, K. Abney, & G. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 55-68). Cambridge, MA: MIT Press.

Altmann, J. (2009). Preventive Arms Control for Uninhabited Military Vehicles. In R. Capurro, & M. Nagenborg, *Ethics and Robotics* (pp. 69-82). Heidelberg: AKA Verlag. Retrieved March 31, 2013, from http://e3.physik.tu-dortmund.de/P&D/Pubs/0909_Ethics_and_Robotics_Altmann.pdf

Arkin, R. (2009). *Governing Lethal Behavior in Autonomous Robots.* Boca Raton: Chapman and Hall/CRC.

Asaro, P. (2008). How Just Could A Robot War Be? In P. Brey, A. Briggle, & K. Waelbers, *Current Issues in Computing and Philosophy* (pp. 50-64). Amsterdam: IOS Press.

Asaro, P. (2011). A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics. In P. Lin, K. Abney, & G. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 169-186). Cambridge, MA: MIT Press.

Beavers, A. (2011). Is Ethics Headed for Moral Behaviorism and Should We Care? *Newsletter on Philosophy and Computers*. Retrieved from http://www.academia.edu/

Borenstein, J. (2008). The Ethics of Autonomous Military Robots. *Studies in Ethics, Law, and Technology 2 (1).* doi:10.2202/1941-6008.1036

Cameron, J. (Director). (1984). *The Terminator* [Motion Picture].

Cashmore, A. (2010). The Lucretian swerve: The biological basis of human. *PNAS 107(10)*, 4499-4504.

Champagne, M., & Tonkens, R. (2012). Bridging the Responsibility Gap in Automated Warfare. *AISB/IACAP World Congress 2012: Symposium on The Machine Question: AI, Ethics and Moral Responsibility* (pp. 67-72). Birmingham, UK: http://events.cs.bham.ac.uk/turing12/proceedings/14.pdf.

Coon, G., & Hamner, R. (1967, February 23). A Taste of Armageddon. *Star Trek: The Original Series.*

Gormley, D. (2008). *Missile Contagion: Cruise Missile Proliferation and the Threat to International Security*. Westport, CT: Praeger Security.

Guarini, M., & Bello, P. (2011). Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters. In P. Lin, K. Abney, & G. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 129-144). Cambridge, MA: MIT Press.

Human Rights Watch. (2012). *Losing Humanity: The Case against Killer Robots.* New York: Human Rights Watch.

ICRAC. (2013, February 27). *The Scientists' Call.* Retrieved March 25, 2013, from International Committee for Robot Arms Control: http://icrac.net/call/

Krishnan, A. (2009). *Killer Robots: Legality and Ethicality of Autonomous Weapons.* Burlington, VT: Ashgate.

Lin, P., Bekey, G., & Abney, K. (2008). *Autonomous military robotics: Risk, ethics, and design.* San Luis Obispo: California Polytechnic State Univ.

Lin, P., Bekey, G., & Abney, K. (2009). Robots in War: Issues of Risk and Ethics. In R. Capurro, & M. Nagenborg, *Ethics and Robotics* (pp. 49-67). Heidelberg: AKA Verlag. Retrieved from http://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1010&context=phil_fac

Lokhorst, G.-J., & van den Hoven, J. (2011). Responsibility for Military Robots. In P. Lin, K. Abney, & G. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 145-156). Cambridge, MA: MIT Press.

Marchant, G., Allenby, B., Arkin, R., Barrett, E., Borenstein, J., Gaudet, L., . . . Silberman, J. (2011). International Governance of Autonomous Military Robots. *The Columbia Science and Technology Law Review, 12*(7), 272-315. Retrieved from http://www.stlr.org/html/volume12/marchant

Minsky, M. (2006). *The Emotion Machine.* New York: Simon & Schuster.

Moor, J. (2006). The nature, importance and difficulty of machine ethics. *IEEE Intelligent Systems, 21(4)*, 18-21.

Saletan, W. (2008, August 11). Ghosts in the Machine: Do remote-control war pilots get combat stress? *Slate*. Retrieved from http://www.slate.com/articles/health_and_science/human_nature/2008/08/ghosts_in_the_machine.html

Sharkey, N. (2011). Killing Made Easy: From Joysticks to Politics. In P. Lin, K. Abney, & G. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 111-128). Cambridge, MA: MIT Press.

Sharkey, N. (2011). The Automation and Proliferation of Military Drones and the Protection of Civilians. *Law, Innovation and Technology 3(2)*, 229-240.

Sharkey, N. (2013, March 27). *Dissecting the Scientists' Call to Ban Autonomous Lethal Robots - Comments.* Retrieved March 29, 2013, from Transhumanity: http://transhumanity.net/articles/entry/dissecting-the-scientists-call-to-ban-autonomous-lethal-robots

Singer, P. W. (2009). *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. New York: Penguin Press.

Sparrow, R. (2009). Predators or Plowshares? Arms Control of Robotic Weapons. *IEEE Technology and Society 28 (1)*, 25-29.

Sparrow, R. (2011). Robotic Weapons and the Future of War. In J. Wolfendale, & P. Tripodi, *New Wars and New Soldiers: Military Ethics in the Contemporary* (pp. 117-133). Burlington, VT: Ashgate.

US Air Force. (2009, May 18). *Unmanned Aircraft Systems Flight Plan 2009-2047.* Retrieved March 25, 2013, from Government Executive: http://www.govexec.com/pdfs/072309kp1.pdf

US DoD. (2012, November 20). *DoD Directive 3000.09.* Retrieved March 25, 2013, from Defense Technical Information Center (DTIC) Online: http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf

Wallach, W., & Allen, C. (2012). Framing robot arms control. *Ethics Inf Technol*, http://dx.doi.org/10.1007/s10676-012-9303-0. doi:10.1007/s10676-012-9303-0

Waser, M. (2012). Safety and Morality Require the Recognition of Self-Improving Machines As Moral/Justice Patients and Agents. *AISB/IACAP World Congress 2012: Symposium on The Machine Question: AI, Ethics and Moral Responsibility* (pp. 92-96). Birmingham, UK: http://events.cs.bham.ac.uk/turing12/proceedings/14.pdf.

Waser, M. (2013, March 27). *Dissecting the Scientists' Call to Ban Autonomous Lethal Robots.* Retrieved March 29, 2013, from Transhumanity: http://transhumanity.net/articles/entry/dissecting-the-scientists-call-to-ban-autonomous-lethal-robots