Ethics Inf Technol (2013) 15:125–135
DOI 10.1007/s10676-012-9303-0

ORIGINAL PAPER

Framing robot arms control

Wendell Wallach · Colin Allen

Published online: 21 October 2012

Abstract The development of autonomous, robotic weaponry is progressing rapidly. Many observers agree that banning the initiation of lethal activity by autonomous weapons is a worthy goal. Some disagree with this goal, on the grounds that robots may equal and exceed the ethical conduct of human soldiers on the battlefield. Those who seek arms-control agreements limiting the use of military robots face practical difficulties. One such difficulty concerns defining the notion of an autonomous action by a robot. Another challenge concerns how to verify and monitor the capabilities of rapidly changing technologies. In this article we describe concepts from our previous work about autonomy and ethics for robots and apply them to military robots and robot arms control. We conclude with a proposal for a first step toward limiting the deployment of autonomous weapons capable of initiating lethal force.

Keywords Military robots · Moral machines · Machine ethics · Operational morality · Robot arms control · Autonomous weapons

Introduction

The development of robotic weaponry is progressing rapidly. The United States and key allies are currently ahead of other countries, but the onset of a robot arms race will eliminate many of the short-term strategic advantages these weapons systems offer. Long-term disadvantages of roboticized warfare are likely to far outweigh the short-term advantages. One of the concerns voiced by critics of military robots is the prospect that robotic weaponry will lower the psychological barriers to starting new wars. Another major concern is that robotic fighting machines in the relatively near future could autonomously initiate lethal activity. Robot arms control has been proposed (Asaro 2008; Borenstein 2008; Altmann 2009; Krishnan 2009; Sparrow 2009, 2011; Sharkey 2011, 2012), but what kinds of prohibitions are likely to gain any traction?

It will be difficult to forge consensus on an appropriate arms-control regime to assuage the first concern. Whether or not robotic weapons would lower psychological barriers to starting wars is, however, immaterial to the need for arms control; arguably, nuclear weapons raised the psychological barrier to war, yet still required international treaties to regulate their use. Given the speculative nature of ideas about the effects of future robotic systems on future warmongering, the demand to regulate robotic weapons is unlikely to compete successfully for the attention of governments facing other arms-control issues that are grounded in the proliferation of more-established technologies.

Among the dangers posed by the second concern, the autonomous initiation of lethal activity, is the unlawful death of non-combatants or friendly forces. Such actions could also start unintended hostilities that quickly escalate. Nevertheless, the use of autonomous systems initiating lethal force has been defended by philosophers including Lokhorst and van den Hoven (2012). However, there is considerable disagreement as to what would constitute an autonomous decision to kill by a robotic system, as well as uncertainty about whether the U.S. military, which

W. Wallach (✉)
Technology and Ethics Research Group, Yale University
Interdisciplinary Center for Bioethics, New Haven, CT, USA
e-mail: wendell.wallach@yale.edu

C. Allen
Department of History and Philosophy of Science, Program in Cognitive Science, Center for the Integrative Study of Animal Behavior, Indiana University, Bloomington, IN, USA

presently holds a technological lead in the development of robotic weaponry, is actually developing autonomous systems that could initiate lethal activity (Singer 2009; Dahm 2012).

Some of the recent discourse uses the phrases "*in*-the-loop" and "*on*-the-loop" (U.S. Department of Defense 2009) to describe the relationship of human supervisors to automated systems. These phrases, however, obscure the degree to which people may or may not be able to intervene in the actions of increasingly autonomous systems. Some authors have proposed that robots may eventually be capable of making moral decisions (Gips 1991; Wallach and Allen 2009) or following the laws of war and rules of engagement (Arkin 2009). These suggestions are sometimes misunderstood as indicating that in the next decade or two robots will have powers of judgment and discrimination that, as a matter of fact, they are unlikely to possess. In an environment where there is almost as much misinformation as information about the capabilities of machines, there is a danger of decisions being delegated to a robotic system based on naive presumptions about its intelligence. In this article we apply concepts from our earlier work to the issues of autonomy in military robots and robot arms control. Our goal is to formulate a specific suggestion for limiting the deployment of potentially dangerous weapons systems, based on a realistic assessment of their capabilities.

Autonomous action by a robot includes any unsupervised activity.[1] This broad definition of autonomy encompasses an array of currently deployed weapons systems, including land mines, cruise missiles, and Aegis and Patriot missile systems. Land mines and cruise missiles are already subject to separate arms-control agreements, and one should expect these to continue to be kept distinct from agreements directed at other kinds of robotic weapons. Nevertheless, the thirty-year attempt to forge arms-control treaties for cruise missiles (Gormley 2008) underscores the challenge of constructing agreements to limit the use of weapons having only low-level autonomy.

Our focus in this paper is on the use of robot weapons that are capable of autonomously initiating lethal activity in such a way that human intervention is practically or technically precluded. The concern here is with weapons whose target acquisition and firing procedures cannot be arrested or are unlikely to be arrested by people either *in* or *on* the loop of decision making. For most of the existing systems humans are "in the loop", which is usually understood as meaning that a responsible human must give a 'go ahead' before the system will initiate lethal activity. The transition from systems where humans are "in the loop" of decision making to "on the loop" was mentioned in a U.S. Air Force report entitled, Unmanned Aircraft Systems Flight Plan 2009–2047 (2009).

> Increasingly humans will no longer be "in the loop" but rather "on the loop" – monitoring the execution of certain decisions. Simultaneously, advances in AI will enable systems to make combat decisions and act within legal and policy constraints without necessarily requiring human input. (U.S. Air Force 2009, p. 41)

In other words, a human monitoring the activity of robots would be limited to vetoing the actions of a system, or a swarm of systems, presuming that the rapid-paced environment of modern warfare allows enough time for the intervention.

We believe that it is of paramount importance to maintain direct human *responsibility* for all actions taken by robotic systems, even if the responsible person does not directly *control* all aspects of system behavior. There is, however, a wide gap between what the robotic systems currently being developed can actually do and anthropomorphic projections of agency to robots, where it is sometimes presumed that robots will soon have "strong artificial intelligence"—the kinds of judgment, sensitivity, and discrimination we expect from wise leaders or good soldiers.

Moral machines

In Wallach and Allen (2009) we map a new field of inquiry that has been variously called Machine Morality, Machine Ethics (ME), Artificial Morality, Computational Ethics, and Friendly AI. This new field of inquiry is directed at the implementation of moral decision-making faculties in artificial agents—i.e., artificial moral agents (AMAs)—and is necessitated by increasingly autonomous systems making choices and taking actions that may cause harm to humans and other subjects of moral concern. The central questions for machine ethics are:

- Do we need artificial moral agents (AMAs)?
  - When? For What?
- Do we want computers making ethical decisions?
- Whose morality or what morality?

[1] The U.S. Army Science Board (2002) describes a scale of ten levels of autonomous behavior beyond Manual—Remote Control (0). At the lowest level is Simple Automation (1) followed by Automated Tasks and Functions (2), Scripted Missions (3), Semi-Automated Missions/Simple Decision Making (4), Complex Missions Specific Reasoning (5), Dynamically Mission Adaptable (6), Synergistic Multi-Mission Reasoning (7), Human-Like Autonomy in a Mixed Team (8), Autonomous Teams with Unmanned Leader/Mission Manager (9), Autonomous—Conglomerate (10). Most of the levels of autonomous behavior are based upon projected future technological capabilities. The discussion in this paper is directed at the development of systems capability of initiating lethal force at level (4) and level (5).

- How can we make ethics computable?

In this article we focus primarily on the significance of the last question for roboticized weapons, and in particular on Ronald Arkin's (2009) proposal 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.

The development of robots capable of functioning as artificial moral agents (AMAs) is likely to be a slow, incremental process. The chart below (Fig. 1), which registers increasing autonomy on its vertical axis and increasing sensitivity to moral considerations on its horizontal axis, will be helpful for appreciating this progression. All technology might be viewed as falling within this chart. A hammer has neither sensitivity nor autonomy. A thermostat has some sensitivity to temperature (an environmental variable that is relevant to human well being) and the autonomy to turn a fan or a furnace on or off when a threshold has been reached. The robotic devices now available or currently being developed are operationally moral in the sense that the designers and engineers who build the system try to anticipate all the kinds of circumstances the robot might encounter and then program an appropriate action for each class of situation. The values instantiated in the robot's choices and actions reflect those of the programmer, and perhaps the broader society, but are also skewed toward those of the corporations who build the device.

As either the environment becomes more complex or the internal processing of the computational system requires the management of a wide array of variables, the designers and engineers who built the system may no longer be able to predict the many circumstances the system will encounter or the manner in which it will process new information. Furthermore, in some situations the information available to the robot will be false, inaccurate, or incomplete. It will thus be necessary for the system to have ethical subroutines through which it determines the safest and most appropriate course of action. Machines with the capacity to explicitly evaluate which of two or more possible courses of behavior is the safest, most acceptable, or most appropriate response to a challenge are functionally moral. Most systems being built today are operationally moral, but increasingly, as the autonomy of robotic systems expands, it will be necessary to create methods for developing functionally moral artificial agents.

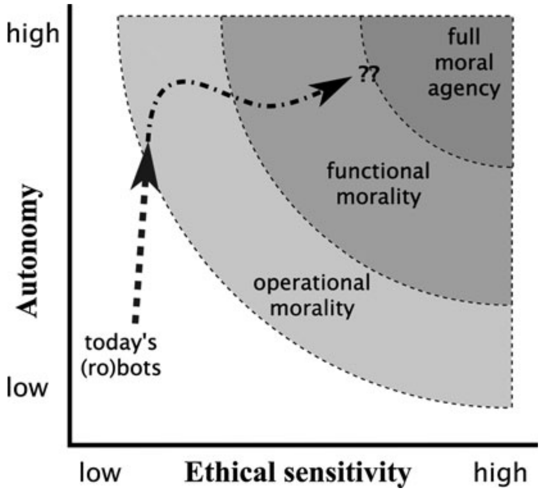

Fig. 1 Two dimensions for the development of artificial moral agents (source hidden for review)

Bounded morality

While it is conceivable that future robots might one day be full moral agents capable of discerning moral responses in a wide variety of contexts and situations, we are far from knowing whether such an eventuality is probable or possible. The behavior of operationally moral or functionally moral robots will, for the foreseeable future, by necessity be bounded. It is possible to design robots that are operationally moral when all the circumstances they will encounter are anticipated in advance. When designers and engineers cannot fully anticipate when and where a functionally moral robot will encounter a challenge they will need to understand:

1. The space (the environment) in which the robot operates well enough to insure that the system recognizes when it is in an ethically significant situation.
2. The routines the system will require for determining an appropriate course of action.

Just as it would be dangerous to put a chain saw in the hands of a child or the hands of an adult who had no training in its use, so too would placing a robot in a context where it would encounter challenges it neither recognized nor had means for determining what actions were safe and appropriate. The bounded morality of a robot will be structured by its intelligence, that is, by its sensitivity to features and changes within that context, and by the ethical routines it has for determining which actions are morally acceptable within that situation. Given that this intelligence will be limited, the environments in which the robot can be safely deployed will also be tightly constrained. If it were possible in a military context to place a robot on a clearly delineated battlefield, with sensors and software that allow it to identify all friendly and unfriendly entities, the challenge of conforming to the rules of war would be simpler than when placing a robot in a partially unknown urban landscape inhabited by both combatants and non-combatants.

Unfortunately, it is naive to presume that robots with bounded morality will only be deployed in contexts for which they are equipped. Even if we can rely on the care of our own military in the deployment of robotic weapons, we have no basis for assuming other parties will demonstrate similar care.

Machine ethics and military robots

The unmanned remotely controlled aircraft (UAVs) and ground vehicles (UGVs) presently favored by the military have limited autonomy and are largely under the control of military supervisors and operators. UAVs and UGVs require a high degree of coordination between the autonomous capabilities of each system and the human operators who direct the system's activity. Deploying the weapons carried by these vehicles usually requires a direct command and an action by the system's operators. In other words, humans are "in the loop" in that no robot will kill without an action by a human agent, who for all practical purposes is directly responsible for the consequences of that action. This position was recently summed up in a U.S. Department of Defense report entitled, "FY2009-2034 Unmanned Systems Integrated Roadmap".

> For a significant period into the future, the decision to pull the trigger or launch a missile from an unmanned system will not be fully automated, but it will remain under the full control of a human operator. Many aspects of the firing sequence will be fully automated but the decision to fire will not likely be fully automated until legal, rules of engagement, and safety concerns have all been thoroughly examined and resolved. (U.S. Department of Defense 2009, p. 10)

> This position is echoed in another report, from the U.S. Air Force:

> 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. (U.S. Air Force 2009, p. 41)

The notion of human supervisors "in the loop" obscures their role in limiting the initiation of kill orders by robots. Consider the RedOwl Sniper Detection Kit which can be mounted on iRobot's PackBot, a military robot which has been deployed by the thousands in Iraq and Afghanistan. Through the use of acoustic direction-finding sensors, high-powered infrared camera, thermal imager, and software the RedOwl can reportedly determine the location of a sniper in an urban setting. Once established, it directs a red laser upon that location as a guide for soldiers to take out the enemy sharpshooter. Arguably, the human soldiers are in the robot's loop rather than vice versa, notwithstanding the fact that there is a military imperative to kill enemy snipers and that soldiers on the ground and robots alike operate under the authority of those in command. Looking ahead to when there is greater acceptance of robots used in combat, mounting the robot with its own armament rather than a laser beam would dispense with the formality of a human to pull the trigger.

Those observing the progression of military robots, including Singer (2009) and Peter Finn of *The Washington Post* (2011) conclude that automated killing drones are already being developed. This idea does not seem far-fetched given that ground-based systems capable of automated killing already exist. SGR-1 sentries, built by Samsung Techwin, and capable of targeting and shooting at large moving objects, are being tested along the Korean Demilitarized Zone (Kim 2010).

Ronald Arkin of Georgia Tech University also believes that lethal autonomous weapons are inevitable. He is directing attention to the development of a method to implement the laws of war, including the Geneva and Hague Conventions, and rules of engagement in robotic systems used in combat. Arkin (2012) believes that if robots outfitted with an ethical governor are provided with the right to refuse unethical orders and the ability to monitor and report unethical behavior of others, the ethical conduct of both human and robotic soldiers will be significantly improved. He bases this belief partially upon a report from the U.S. Surgeon General's Office, Mental Health Advisory Team (MHAT) IV (U.S. Army Medical Department 2008) that underscores the ethical failings of soldiers in warfare. Among MHAT's findings: approximately 10 % of soldiers and Marines report mistreating non-combatants, only 47 % of soldiers and 38 % of Marines agreed that non-combatants should be treated with dignity and respect, and only 45 % of soldiers and 60 % of Marines would report a team member for unethical behavior. There are many understandable reasons why human soldiers and Marines would fail to follow ethical guidelines, but these reasons do not excuse behavior that violates military codes of conduct. Arkin (2009) theorizes that robots would be more humane than human beings in military situations and will improve the overall conduct of war. He gives several reasons for this view: robot fighting machines can be designed without emotions such as anger or the desire for revenge; through their sensors and networks they may have access to vast quantities of information; their computing power could give them the capacity to integrate more information than a human soldier would; and sharing the battlefield with human soldiers they could monitor the humans' behavior. To be sure, Arkin has no illusions that robots will be ethically perfect on the battlefield, only that they can be designed to perform more ethically than humans do.

Many technical challenges must be surmounted in order to implement the laws of war and rules of engagement in a computational system. Even if successful, Arkin has been quite clear that systems outfitted with an ethical governor[2] should only be used for circumstances in which systems with bounded morality might be able to effectively cope—discrete specialized missions where the scope is limited to tasks such as, room clearing, counter-sniper operations, and perimeter protection. Arkin has agreed that such systems are not appropriate for the full range of counterinsurgency operations and that they should only be deployed when the likelihood of encountering civilians has been minimized. He is well aware of some of the inherent limitations of the robotic systems that will be available within the next decade or two. However, others, hearing his claims that robots will be ethical soldiers may be less cognizant of those limitations.

We do not share Arkin's presumption that lethal autonomy is inevitable, and we believe that efforts should be made to prevent this possibility. However, we have also been supportive of implementing moral decision making faculties in robots. Given that we could be wrong about the inevitability of lethal autonomy, we cautiously applaud his continuing efforts, and appreciate that someone is thinking about this challenge. Nevertheless, we are concerned that Arkin's proposal might lead less informed parties to underestimate the difficulty of the task, overestimate the capacities of the machines, and therefore fail to understand the limited ways in which such systems should be deployed. The belief that the laws of war and rules of engagement can be implemented could spur the development of lethal autonomous weapons systems based upon an unproven thesis that adequate self-restraint can be built into such systems. And even if moral decision making capability can be built into military robots, only a few technologically sophisticated countries are likely to have systems with such capabilities. Confronted with asymmetrical warfare other state and non-state actors may well turn to lethal autonomy with little or no restraint built into their systems. For these reasons we support an outright prohibition on the deployment of autonomously mobile lethal robots, and make specific proposal to this end further below. Before providing that proposal, however, we survey in greater detail the challenges and risks confronting anyone who would pursue the path that Arkin has chosen.

AMAs: two hard problems

The task of building AI systems with even modest moral decision-making capabilities faces two hard problems. The first requires selecting a set of norms, rules, principles, or procedures for the system to use in making moral judgments, and finding a computational method to implement them. Most moral philosophers appreciate that finding an ethical theory or rules to cover all cases adequately is itself a daunting, if not impossible, feat. Some philosophers argue from a theoretical position that morality is not the kind of thing that can be implemented in a machine (e.g., Stahl 2002) while others have argued that the project of reasoning ethically from general principles is misguided (e.g., Dancy 2011). Rather than implementing very general principles, Arkin proposes to implement the laws of war and specific rules of engagement for a particular conflict, but it is far too early to assess whether the strategy he has selected for the design of an ethical governor for military robots would be effective in the situations for which it is intended. Even if one finds a computational architecture that is theoretically adequate for ethical decision making in the battlefield, another, more important step will be to test whether such a system selects appropriate and acceptable actions in real world situations.

The second hard problem concerns how to set boundaries to the assessments that must be carried out for effective moral decision making. This is actually a group of related challenges. How does the system recognize that it is in an ethically significant situation? How does it discern essential from inessential information? How does the AMA estimate the sufficiency of initial information? What capabilities would an AMA require to make a valid judgment about a complex situation, e.g., combatants vs. non-combatants? How would the system recognize that it had applied all necessary considerations to the challenge at hand or completed its determination of the appropriate action to take? For example, what stopping procedure would the system use to determine that it had completed a utilitarian calculation?

To be sure, humans can sometimes fail in all of the ways implicit in these questions, for instance making mistakes in tasks such as determining who is a combatant and who is a non-combatant. Nevertheless, humans bring powers of discrimination to bear to performing such tasks that we either do not know how to implement in robots, or for which we have at best a few rudimentary theories.

This group of challenges is related to the frame problem, both as it was first elucidated by AI researchers (McCarthy and Hayes 1969) and as it was later embellished by philosophers (Dennett 1978; Fodor 1983) to cover wider epistemological issues. In AI, the frame problem concerns how to represent only those effects of an action that are relevant to choosing among actions without having to also explicitly represent all the intuitively obvious mundane effects. For philosophers the problem extended to how any intelligent agent would limit the set of beliefs that must be re-evaluated and possibly changed as the result of an action.

[2] Arkin does propose that a human supervisor could override the ethical governor.

Frame problems arise in implementing any norms, rules, principles, or procedures in an AMA. An AMA functioning in anything other than a tightly bounded context will carry a heavy computational load as it will need to estimate the sufficiency of the initial information available and search out sources for additional information, it will be required to have significant psychological knowledge about the other actors in the environment, and it will need to have knowledge of effects of actions (its own and that of other actors) in the world. The difficulty for an AMA is that the boundaries for evaluation of its possible actions are potentially unlimited. Nevertheless, humans manage to function with a number of heuristic and affective processes that effectively limit the kinds of unlimited search that seem to be demanded by more formal procedures. How to implement these in AI is unclear, a point to which we return below.

Norms, rules, and principles

Asimov's laws are what first come to mind for many people when they consider ways to constrain a robot's behavior. The three laws, and a Zeroth Law that Asimov added later, are:

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

> Zeroth law: a robot may not injure humanity, or, through inaction, allow humanity to come to harm.

Asimov was writing fiction, he was not building robots. However, in story after story he illustrated how a robot equipped with intuitively straightforward rules arranged hierarchically would fail. For example, what should a moral machine programmed with the three laws do if it receives conflicting orders from two different humans? Asimov's stories illustrate that rules alone will not be adequate to insure moral behavior from a robotic system.

Robots capable of even limited autonomous activity will need to factor in an array of considerations in determining what behavior is appropriate or legal when confronted with difficult ethical challenges. The field of machine morality is largely concerned with the approaches and procedures used by the robot to make such judgments. We have written about this subject extensively in *Moral Machines* (2009), and so we will only mention a few brief details here.

The approaches for implementing moral decision-making capabilities in robots fall within two broad categories, top-down and bottom-up. Top-down refers to the implementation of rules, principles or moral decision-making procedures, such as utilitarianism, Kant's categorical imperative, the Ten Commandments, Hinduism's yama and niyama, and Asimov's laws. A top-down approach takes an antecedently specified ethical theory, or a set of context-specific principles or rules, such as an ethical code, and analyses the requirements for computational implementation. Arkin's (2009) architecture is an example of a top-down approach. Bottom-up approaches take their inspiration from theories of learning, evolutionary psychology and game theory, as well as developmental psychology and theories of moral development. Bottom-up approaches, if they use a prior theory at all, do so only as a way of specifying the task for the system, but not as a way of specifying an implementation method or control structure.

Both top-down and bottom-up approaches have strengths and weaknesses. For example, it is a strength of principles that they may be defined broadly to cover countless situations, but a weakness of being broad or abstract is that their application to specific situations will be debatable. Bottom-up approaches are particularly good at dynamically integrating input from discrete subsystems. But defining the ethical goal for a bottom-up system would be difficult, as would assembling a large number of discrete components into a functional whole.

Eventually, more complex applications may need AMAs that maintain the dynamic and flexible morality of bottom-up systems, to accommodate diverse inputs, while subjecting the evaluation of choices and actions to top-down principles, to represent ideals people strive to meet.

Beyond reason

We have argued (Wallach and Allen 2009) that a broad range of situations require AMAs to have capabilities in addition to the ability to reason. These supra-rational capabilities (beyond reason) include emotions, social intelligence, empathy, a theory of mind, consciousness, and being embodied and embedded in a world with humans, objects, and other agents. These capabilities may serve a broad range of purposes, including allowing access to information and providing input to decisions without the need for all the relevant facts and knowledge to be formally and explicitly represented. There is currently only rudimentary understanding of how these function in humans. Nevertheless, computer scientists have already initiated new fields of research to instantiate functional equivalents of emotions, theory of mind, and consciousness within computers and robots. There are many questions about the feasibility and necessity of implementing all these capabilities. But one of the tasks for machine ethics is to delineate the capabilities AMAs will require in order to operate appropriately and safely within specific domains.

Of particular interest are emotions and their role in decision making. The Stoic philosophers saw emotions as a hindrance to moral reflection, but with the onset of research on emotional intelligence there has been considerable reevaluation of the Stoic position. Although Arkin initially suggested that emotionless robots are less likely to commit moral atrocities than human soldiers, he has more recently conceded that some limited emotional capacities may be necessary. He notes a possible need for (2009, p. 140), "the use of a strict subset of affective components, those that are specifically considered the moral emotions". He goes on to write (p. 140) that, "an architectural design component modeling a subset of these affective components (initially only guilt) is intended to provide an adaptive learning function for the autonomous system architecture should it act in error." Emotions other than guilt (e.g., fear) may also play important learning functions. The possibility of gaining the adaptive benefits of emotions by way of their cognitive representation is yet to be proved.

Risks in developing AMAs for combat

In the development of AMAs, what kinds of risks are acceptable? Given the difficulties inherent to implementing moral decision making within computational systems, is it wise to be developing AMAs for realms like warfare before we have demonstrated a proof of concept in systems developed for non-military applications? Noel Sharkey (2012) has been pointing out for years that robots do not now, nor will they soon, have the ability to discriminate between friend and foe, and until they do we should not even begin to consider allowing them to make 'decisions' on the battlefield. Andreas Matthias (2011) argues that the kinds of calculative systems of morality such as the ethical governor Arkin proposes:

> are in principle only able to deal with a conflict-free subset of rule-based ethics, since they lack all mechanisms which are commonly assumed to be necessary for resolving moral rule conflict: phronesis, moral intuitions, or an understanding of human preferences and the utilitarian value of specific consequences to each affected person. But this 'toy ethics' is not sufficient to resolve real-world moral problems on the battlefield, which typically involve conflicting options about questions of life and death, of justified causes, of retribution and retaliation, and of culture-specific ethics codes. (Matthias, 2011, p. 300)

Matthias, like Sharkey, is asserting that the systems proposed by Arkin have fewer capabilities and are thus more risky than Arkin acknowledges. We concur. Nevertheless, there will be ready financing for military applications, which would make it possible to conduct extensive risk assessment. We think there is a risk that such systems could be deployed without sufficient verification and testing, especially in countries with fewer resources. We doubt that Arkin would disagree that such a risk exists, but he and military planners in many countries would argue that sufficient testing regimes are in place, at least in the more developed countries. Nevertheless, despite best intentions, it is very hard to be sure that one has ever conducted all the tests that would be needed for a high degree of certainty about the nature of the risks involved. And under wartime conditions, weapons systems may be deployed before they are fully tested. This has been the case with some of the UAVs deployed by the U.S. in Iraq, Pakistan, and Somalia (Department of Defense Task Force Report 2012). Therefore we see no ultimate resolution to the dispute between those who think that the implied level of risk is acceptable and those who think it is not. Given that some key players will be willing to accept some level of risk associated with the use of robots used in war, in the end the risk they pose is secondary to the ethical question of whether in principle robots ought to be engaged in lethal autonomy.

The illusion of full autonomy

Scholars in the field of Cognitive Systems Engineering study the interaction between workers and the technologies they use. For all practical purposes, with the exception of a few limited purpose machines, an intelligent system and the operators who work with it are best understood as a Joint Cognitive System (JCS). JCSs require tight coordination between the activities of the human and the mechanical component. Given that the actions of the mechanical components within a JCS tend to be limited, increased complexity of the system usually results in additional demands on the flexibility of the human operators.

Woods and Hollnagel (2006) note that with the advent of artificial agents, when a JCS fails there is a tendency to blame the human as the weak link, and to propose increased autonomy for the mechanical device as a solution. Furthermore, there is the illusion that increasing autonomy will allow the designers to escape responsibility for the actions of artificial agents. However, Woods and Hollnagel point out that increasing autonomy of the artificial components will actually add to the burden of human operators. They illustrate this with the example of an accident on December 6th, 1999, that caused $5.3 million in damages when there was a failure in coordination between operators and a semi-autonomous Global Hawk UAV. Maneuvering the Global Hawk on the ground, the operators misunderstood the system's actions. The conflict between what the system was doing and what the operators thought the system was doing led to the aircraft going off the runway, where its nose gear collapsed. In a more recent event, Iran captured a CIA Drone, and while the details surrounding this incident are unclear, indications that the operator of the UAV lost control of the vehicle demonstrate once again how difficult it can be to manage a JCS.

The behavior of robots will continue to be brittle on the margins as they encounter new or surprising challenges. Human operators will need to anticipate what the robot will try to do in new situations in order to effectively coordinate their actions with those of the robot. However, anticipating the robot's actions will be harder to do as systems become more complex and independent, leading to a potential increase in conflicts between the actions initiated by the system and the actions initiated by the human operators. While each failure may be attributed to the operators, to expect operators to anticipate the actions of intelligent systems becomes more and more unreasonable as the systems and the environments in which they operate become more complex.

Managing complex adaptive systems

There are inherent and perhaps inevitable problems in managing complex adaptive systems. As mentioned above, it becomes harder for humans to coordinate with artificial systems when they become more independent, and the likelihood of 'black swans'—incidents which have a low probability of occurring but if they occur will have a high impact—increases (Taleb 2007).

Following the analysis of Woods and Hollnagel (2006), Hollnagel et al. (2006), complex adaptive systems fail when:

1. The autonomous adaptive system exhausts its capacity to adapt as disturbances/challenges cascade (decompensation).
2. The operators and/or the autonomous system exhibit behavior that is locally adaptive but globally maladaptive leading to the initiation of actions that work at cross-purposes.
3. When the operators and/or the complex adaptive system is stuck in outdated behavior, or there is an over-reliance on behavior that was successful in the past.

They propose that designers and engineers should strive to engineer greater resilience into JCSs. But there may be fundamental limits on how successful this endeavor will be. Not all of the challenges posed by complex systems can be ameliorated with more attention in the design of JCSs to coordination between the human operators and the intelligent components.

Monitoring, assessing, and verifying: arms control specific issues

As we stated earlier, the development of AMAs is likely to be a long, incremental process. Throughout this development, a primary challenge for society will be monitoring and assessing the capabilities of each system. What criteria should be used to determine whether a particular system could be deployed safely in a specific context? What oversight mechanisms need to be put into place in order to ensure that such an assessment can be made and has been made? What penalties might be applied if a certified system is later implicated in harmful actions? In principle, we believe that decisions to initiate lethal force should never be delegated to autonomous systems. However, the range of military systems in which the term "autonomous" can be applied is so wide as to make a blanket ban on all autonomous weapons highly unlikely. To attempt to govern everything from automatic sentry systems deployed at contended borders to long-range drones flying in foreign airspace under a single treaty would introduce so many complexities as to be unworkable. For instance, the new capabilities that derive from the integration of GPS with legged robots capable of traveling over almost any kind of terrain provides a kind of autonomous mobility that could allow an explosive payload to be delivered, without direct human oversight, to places currently out of reach by either wheeled vehicles or air strikes. Should such robots, which provide new military capabilities, be banned outright, banned only from carrying certain kinds of weapons, or banned from carrying weapons of any kind so that they may be used only for tactically legitimate surveillance purposes? Given the flexibility with which software and hardware components may be recombined for different kinds of applications, could a ban on one application of the technology provide safeguards against inappropriate use for other applications (Lin 2011)? Because the considerations for different kinds of military robots diverge so much, we think that autonomy will have to be considered in the context of developing arms control agreements for specific weapons systems, rather than treating autonomous systems per se as the focus of a single arms control treaty.

Even within the framework of weapons-specific agreements, autonomy raises a number of difficult questions. Verification has always been an almost insurmountable issue for arms control, but presuming it were not: How could one verify the capacity for human-independent activity by a weapons system? Would an arms-control regime need access to hardware, software, operating manuals and field tests of autonomous operations? The possibilities for hiding the full operational capabilities of sophisticated devices seem much greater than were afforded in the days when the game was primarily one of hiding weapons from enemy view while perhaps downplaying the full extent of their payload, physical range, and the associated telemetry and targeting capabilities of those supporting these systems. When limitations in capacities for autonomous activity are more due to software than physics, the upgrade paths may be more rapid and less detectable, for the same physical form can embody very different capacities. And while autonomous targeting might need to be verified by arms control inspectors separately from autonomous firing, it is a very small change in the design to go from having acquired a target automatically to pulling the trigger. Thus inspections may not create a large impediment to deployment of systems that combine these capabilities.

Furthermore, if discrimination and moral decision making capabilities are touted as providing ethical governance for specific battlefield systems, these capabilities will also need to be verified and tested by international inspectors in a variety of contexts. The appropriate deployment of systems with limited autonomy for specific contexts will vary as their sensitivity to moral considerations vary. In the past, weapons inspectors have not been concerned with every bug fix and software upgrade to existing weaponry. But once the autonomy of killing machines becomes a concern, the frequency and likelihood of software upgrades would require an unmanageable regime of constant reverification. Given that any form of initial verification will be difficult if not impossible to agree upon, this alone could be an insurmountable hurdle.

These points will not be new to arms control experts, who have long struggled with the fact that as technology changes, the boundaries between supposedly different kinds of weapon systems collapse (Gormley 2008). Arms control agreements tend to be retrospectively focused on systems that are entirely within existing technological capabilities (although there are exceptions such as the ban on biological weapons that covers all potential future biological agents through a general-purpose criterion), rather than on what might not even be technically feasible. Unfortunately, waiting until after many countries and private corporations have developed autonomous lethal weapons systems will lead to the alignment of many vested interests against arms control that limits their deployment.

An initial step

Autonomous weaponry capable of initiating lethal force may be considered a violation of international humanitarian law, although this point has not been clarified by any international authority. In addition, a few U.S. military leaders are privately beginning to express concern about the loss of robust command and control posed by future autonomous systems capable of initiating lethal force. These two points suggest that there is a unique opportunity to prospectively limit weaponry capable of autonomously initiating lethal activity.

Loss of command and control could lead to:

- Unintended initiation of hostilities.
- Collateral damage—increase in collateral damage downstream.
- Failure of missions arising from the poor coordination between soldiers and their semi-autonomous weapons, and risk of friendly-fire casualties.
- Cultural backlash, particularly in counter-insurgency operations where managing relations with the local populace is critical to success.
- Deployment of autonomous lethal force by other governments and non-governmental actors.
- Future wars pitting autonomous lethal weaponry against each other.
- Potential for destabilizing strategic military balances.

As a first step, we propose an executive order from the President of the United States that clarifies limits on initiation of lethal activity by the autonomous weapons systems (UAVs, UGVs, and UWVs) that the U.S. will deploy. We offer three possible courses of executive action:

1. Declaration that a deliberate attack with lethal force by fully autonomous weaponry violates the laws of war. The executive order would establish that this principle already exists in international law.
   - Advantage: This strategy affirms the U.S.'s commitment to the rule of law and seeks to clarify a critical international humanitarian principle that would also protect the U.S. from such attacks.
   - Disadvantage: Precludes developing certain types of weapons in the future.
2. Declaration that the U.S. will not deploy such weaponry.
   - Advantage: Keeps open the possibility of overturning this principle in the future, if warranted.
   - Disadvantage: Weakens the principle, and is unlikely to deter governments and corporations from developing autonomous lethal force, especially if the U.S. and U.S. corporations continue to develop such weapons and others perceive this as a strategic threat.
3. Declaration that the U.S. will observe a 10-year moratorium on developing and deploying such weapons, in order to assess the ethics and legality of the

systems as well as build international consensus on this issue.

- Advantage: Ability to develop such weapons in the future if there is no international agreement in the meantime that deters their development.
- Disadvantage: Will have little effect on deterring the development of such weapons.

A central challenge will lie in defining the class of weapons covered by such an order. One possible formula is to define the class as offensive systems that could select targets and initiate lethal activity.[3] The declaration would state that a human must be kept "in the loop" for at least one of these two activities.

We recognize that this proposal is U.S. centric and does not speak directly to the challenge that many nations are developing robotic weaponry. A unilateral declaration by the U.S. will not in itself stop the development of autonomous lethal force. However, given the strategic advantage in unmanned systems currently held by the U.S., and the difficulty in getting the U.S. to sign arms control agreements, a move by the U.S. to establish limits based upon humanitarian concerns will carry some moral force. That moral force will be compounded if NATO follows suit. This could in turn establish a principle under international humanitarian law. Such an approach could avoid laborious arms-control negotiations on details of verification and inspection.[4]

Any declaration by the U.S. and its NATO allies would only be a first step, and would need to be followed up with commitments from other members of the international community to forego autonomous lethality. If the international community cannot agree to forego the development of autonomous lethal weaponry, autonomy will need to become a consideration in all future arms control negotiations. For each context, specific issues of autonomous operation can be addressed and we urge that all present arms treaties be re-evaluated with such questions in mind. In the same way that range and payload are presently part of the standard vocabulary for arms agreements, autonomy should be considered essential to all future negotiations.

The development of autonomous weaponry is truly a game changer. Indeed, given the manner in which autonomous systems could radically alter the conduct of future wars, it would be advisable to promote a worldwide adoption of the principle that robots should not initiate lethal activity, even while the application of this principle to specific systems will be the subject for future negotiations.

[3] The U.S. military will wish to permit autonomous defensive systems such as anti-ballistic missile systems (e.g., Patriot) and ship defense systems (e.g., Phalanx).

[4] We are grateful to Jürgen Altmann and an anonymous reviewer for pointing this out.

References

Altmann, J. (2009). Preventive arms control for uninhabited military vehicles. In R. Capurro & M. Nagenborg (Eds.), *Ethics for robotics*. AKA Verlag, Heidelberg.

Arkin, R. (2009). *Governing lethal behavior in autonomous robots*. Chapman and Hall: CRC.

Arkin, R. (2012). *Presentations at the EPIIC international symposium on conflict in the 21st century*. Tufts University, February 22, 23.

Asaro, P. (2008). How just could a robot war be? In P. Brey, A. Briggle, & K. Waelbers (Eds.), *Current issues in computing and philosophy* (pp. 50–64). Amsterdam, The Netherlands: IOS Press.

Borenstein, J. (2008). The ethics of autonomous military robots. *Studies in Ethics, Law, and Technology 2*(1): Article 2. doi: 10.2202/1941-6008.1036. Available at: http://www.bepress.com/selt/vol2/iss1/art2.

Dahm, W. J. A. (2012). Killer robots are science fiction. *The Wall Street Journal*, February 16th 2011. Available online at http://online.wsj.com/article/SB10001424052970204883304577221590015475180.html. Accessed 13 Oct 2012

Dancy, J. (2011). Contribution to discussion on "The Future of Moral Machines", *On the human.* National Humanities Center. http://onthehuman.org/2011/12/the-future-of-moral-machines/. Accessed 1 May 2012.

Dennett, D. C. (1978). *Brainstorms*. Cambridge: MIT Press.

Finn, P. (2011). A future for drones: Automated killing. *The Washington post,* September 19, 2011. Available online at http://www.washingtonpost.com/national/national-security/a-future-for-drones-automated-killing/2011/09/15/gIQAVy9mgK_story.html. Accessed 19 December 2011.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge: MIT Press.

Gips, J. (1991). Towards the ethical robot. In K. G. Ford, C. Glymour, & P. J. Hayes (Eds.), *Android epistemology* (pp. 243–252). Cambridge: MIT press.

Gormley, D. M. (2008). *Missile contagion: Cruise missile proliferation and the threat to international security*. London: Praeger.

Hollnagel, E., Woods, D. D., & Leveson, N. (Eds.). (2006). *Resiliance engineering: Concepts and precepts.* Aldershot: Ashgate Publishing.

Kim, T.-G. (2010). Machine gun-armed robots to guard DMZ. *The Korea Times,* June 24, 2010. Available online at http://www.koreatimes.co.kr/www/news/biz/2010/06/123_68227.html. Accessed 19 December 2011.

Krishnan, A. (2009). *Killer robots: Legality and ethicality of autonomous weapons*. Burlington: Ashgate.

Lin, P. (2011). Drone-ethics briefing: What a leading robot expert told the CIA, *The Altantic*. December 15, 2011. Available online at http://www.theatlantic.com/technology/archive/2011/12/drone-ethics-briefing-what-a-leading-robot-expert-told-the-cia/250060/. Accessed 19 December 2011.

Lokhorst, G., & van den Hoven, J. (2012). Responsibility for military robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics.* Cambridge: MIT Press.

Matthias, A. (2011). Algorithmic moral control of war robots: Philosophical questions. *Law, Innovation and Technology, 3*(2), 279–301.

McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In D. Michie, & B. Meltzer (Eds.), *Machine Intelligence 4* (pp. 463–502). Edinburgh: Edinburgh University 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. (2012). Killing made easy: From joysticks to politics. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics*. Cambridge: MIT Press.

Singer, P. W. (2009). *Wired for war*. 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 (Eds.), *New wars and new soldiers: Military ethics in the contemporary world* (pp. 117–133). Surrey, UK & Burlington, VA: Ashgate.

Stahl, B. C. (2002). *Can a computer adhere to the categorical imperative? A contemplation of the limits of transcendental ethics in IT*. Paper presented at the international conference on systems research, informatics and cybernetics, Baden Baden, GE.

Taleb, N. N. (2007). *The black swan: the impact of the highly improbable*. New York: Random House.

U.S. Army Science Board. (2002). Ad Hoc study on human robot interface issues. Available online at http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA411834. Accessed 19 April 2012.

U.S. Army Medical Department. (2008). *MHAT-IV*. http://www.armymedicine.army.mil/reports/mhat/mhat_iv/mhat-iv.cfm. Accessed 20 December 2011.

U.S. Air Force. (2009). *Unmanned Aircraft Systems Flight Plan 2009–2047*. Available at http://www.govexec.com/pdfs/072309kp1.pdf. Accessed 19 April 2012.

U.S. Department of Defense (2009). *Fiscal Year 2009–2034 Unmanned systems integrated roadmap*. http://www.acq.osd.mil/psa/docs/UMSIntegratedRoadmap2009.pdf. Accessed 20 December 2011.

U.S. Department of Defense. (2012). *Task force report: The role of autonomy in DoD systems*. http://www.fas.org/irp/agency/dod/dsb/autonomy.pdf. Accessed 22 September 2012.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Woods, D. D., & Hollnagel, E. (2006). Joint cognitive systems: Patterns in cognitive systems engineering. Boca Raton: CRC Press.