Ethics Inf Technol (2014) 16:51–62
DOI 10.1007/s10676-013-9335-0

ORIGINAL PAPER

Negotiating autonomy and responsibility in military robots

Merel Noorman · Deborah G. Johnson

Published online: 18 February 2014

Abstract Central to the ethical concerns raised by the prospect of increasingly autonomous military robots are issues of responsibility. In this paper we examine different conceptions of autonomy within the discourse on these robots to bring into focus what is at stake when it comes to the autonomous nature of military robots. We argue that due to the metaphorical use of the concept of autonomy, the autonomy of robots is often treated as a black box in discussions about autonomous military robots. When the black box is opened up and we see how autonomy is understood and 'made' by those involved in the design and development of robots, the responsibility questions change significantly.

Keywords Autonomy · Responsibility · Military robots

Introduction

Drones are now commonplace in American military engagements, and many predict that more sophisticated and more autonomous robots will increasingly be incorporated into military operations on the battlefield (see for example Adams 2001; Lin et al. 2008; Singer 2009; U.S. Air Force 2010). Central to the ethical concerns raised by the use of autonomous military robots are issues of responsibility and accountability. Who will be responsible when these technologies decide for themselves and behave in unpredictable ways or in ways that their human partners do not understand? For example, if an autonomously operating, unmanned aircraft crosses a border without authorization or erroneously identifies a friendly aircraft as a target and shoots it down (Asaro 2008)? Will a day come when robots themselves are considered responsible for their actions (Lin et al. 2008)?

Literature in the field of Science and Technology Studies (STS) shows that the trajectory of technological development is contingent, multidirectional, and dependent on complex negotiations among relevant social groups (Bijker et al. 1987). Technologies that are adopted and used are not predetermined by nature or any other factor, and cannot be predicted in advance with certainty. In the course of development, the design of a new technology may morph and change in response to many factors including changes in funding, historical events such as wars, changes in the regulatory environment, accidents, market indicators, etc. The technologies that succeed (i.e., are adopted and used) are the outcome of complex negotiations among many different actors including engineers and scientists, users, manufacturers, the public, policy makers, and others.

Negotiations among the actors relevant to a new technology are reflected, and can be observed, in the discourse around the new technology in its earliest stages of development. The discourse around responsibility and autonomous military robots is a case in point; current discourse provides an opportunity to observe issues of responsibility being worked out. The negotiations between researchers, developers, engineers, philosophers, policy-makers, military authorities, lawyers, journalists, human-rights activists, etc. are taking place in the media and academic

M. Noorman (✉) · D. G. Johnson
University of Virginia, Charlottesville, VA, USA
e-mail: merel.noorman@ehumanities.knaw.nl

D. G. Johnson
e-mail: dgj7p@virginia.edu

M. Noorman
eHumanities, Royal Netherlands Academy of Arts and Sciences,
Amsterdam, The Netherlands

journals, at conferences and trade shows, through drafting of new policies and regulations, in negotiating international treaties, and, also, of course, through designing and developing the technologies. This sharply contrasts with the all too common idea that issues of responsibility are predetermined or decided separately from technological design or after a technology is developed.

At the broadest levels of analysis, the responsibility issues posed by the prospect of increasingly autonomous military robots are not difficult to grasp. In moral philosophy and more generally, autonomy implies acting on one's own, controlling one's self, and being responsible for one's actions. Being responsible for one's action in particular requires that the person had some kind of control over the outcome at issue. Thus, framing robots as becoming increasingly autonomous may suggest that robots will be in control and that human actors will, therefore, not be in control. Hence, humans cannot be held responsible for the behavior of autonomous robots.

However, in this paper, we argue that more machine autonomy does not necessarily mean less human responsibility. In order to understand the issues of responsibility we have to look more closely at the relationship between human control and machine autonomy. That requires a closer analysis of the various conceptions of machine autonomy that now pervade the discourse on autonomous military robots. We are not concerned with whether robots are or are not autonomous. Rather we are interested in how different conceptions of machine autonomy interact with questions of responsibility. We therefore focus on the ways in which different actors use and conceive of autonomy to characterize robotic systems that are currently in the early stages of development. Our argument is that leaving the notion of autonomous robots unspecified or underspecified draws attention away from important choices that are made at the level of design and implementation, choices that structure the possibilities for responsibility.

Negotiating responsibility for autonomous robots

The discourse around responsibility and autonomous robots is rich and complex. When it comes to the question of who will be responsible for the behavior of autonomous robots, especially autonomous robots of the future, a variety of positions have been articulated and a struggle over which is likely to, or should, be adopted can be observed in the discourse. As already suggested, some argue that it will not be possible to hold humans responsible for the behavior of autonomous robots (Matthias 2004; Sparrow 2007). Others entertain the idea that autonomous robots might someday be held responsible in some narrow sense for their own behavior (Hellstrom 2012; Asaro 2007; Wallach 2013).[1] Some argue that responsibility will or should be shared between robots and the human actors involved (Crnkovic and Persson 2008; Crnkovic and Çürüklü 2012). Yet others argue that humans will always be responsible for the behavior of robots (Nagenborg et al. 2008; Marino and Tamburrini 2006; Chopra and White 2011). Those who take this position sometimes argue that there is nothing new about autonomous robots in the sense that the legal and moral concepts we currently use for other complex technologies will be adequate to handle responsibility for autonomous robots.

Perhaps the most striking feature of the discourse on responsibility and autonomous robots (as compared to the discourse of engineers, computer scientists and designers discussed in the next section) is the extent to which the autonomy of robots is underspecified and left as a black box. Broadly, autonomy is used in this discourse to refer to machines that require no human intervention for their operation. Wallach and Allen (2013) recently specified autonomous action by a robot simply as "unsupervised activity". In the discourse, the claim that robots have autonomy and will continue to have more is often used as a premise—a starting place—for thinking about the implications of such robots. This means that little attention is paid to what goes on inside the robots.[2] The argument of this paper is that attention to what goes on inside robots is essential to understanding the responsibility issues.

To be sure, there are reasons for underspecifying what is meant by autonomous robots. For one, the discourse is focused on the future and we can only speculate about how future robots will operate. Participants in the discourse use concepts like autonomy, learning, and decision-making metaphorically to characterize the envisioned robotic systems as having abilities comparable to familiar human abilities. We have an idea of what it means to be autonomous in the case of human beings and these future narratives extend that idea to machines. The use of such metaphorical concepts may then suggest that the notion of increasingly autonomous robots requires little further explanation beyond referring to the corresponding human capacities.

[1] Hellstrom argues not exactly that autonomous robots will be responsible but that we will be inclined to consider them responsible when they are responsive to praise and blame. Asaro (2007) entertains the possibility of robots being legally liable and subject to punishment by comparing legal liability for robots to the legal liability of corporations. Wallach (2013) suggests that: "If and when robots become ethical actors that can be held responsible for their actions, we can then begin debating whether they are no longer machines and are deserving of some form of personhood."

[2] There are exceptions to this as in the case of Matthias (2004) who specifies several different kinds of programming that are considered autonomous.

Yet, another reason for underspecifying is that autonomy is not a clear or well-understood concept (Lee and Brown 1994). Even in moral philosophy, where human autonomy underpins the very possibility of ethics, autonomy is a highly contested concept. Autonomy is what distinguishes humans from other kinds of entities and what makes morality possible, but what it is and how humans can have it and still be subject to the deterministic world continues to be subject to debate in various philosophical traditions, including action theory and metaphysics. The variable meaning of the concept seems to facilitate agreement that: robots will become more autonomous; that enhanced autonomy will make robots incomprehensible to humans; and, therefore, that humans will not be able to control them.

The variable meaning of the concept of autonomy can lead to miscommunication and unjustified expectations. Metaphorical uses of concepts, like autonomy, are instrumental in allowing us to talk about and refer to things that do not yet exist, but they are open to interpretation and the similarities that one person draws between human and machine autonomy might be different from the ones that another person draws. For instance, moral philosophers tend to have a different conception of autonomy as compared to computer scientists (Noorman 2009). We therefore have to look closely at the various conceptions of autonomy that are being used in discourse on autonomous robots.

The consequences of leaving the notion of autonomous robots underspecified can be illustrated by examining two important positions in the discourse on responsibility and autonomous robots: Sparrow's argument (2007) against killer robots and the suggestions that robots might be held responsible for their own behavior.

No one will be responsible

Sparrow's argument in "Killer Robots" (2007) provides an illustration of how an underspecified notion of autonomy is used to justify the claim that no humans will be responsible for the behavior of (future) autonomous robots. For Sparrow the claim that no humans can be responsible leads to a focus on whether autonomous military robots should ever be built or used. In this paper, our concern is not to disagree with Sparrow's claim that it is unethical to use unpredictable robots for which no human can be responsible in armed conflicts. Our concern is to show how underspecified notions of autonomy direct attention away from the range of possibilities for future robots and in particular how the design of robots makes a difference in whether and how humans can be responsible for them.

Sparrow argues that the use of autonomous military robots is unethical because no humans can be held responsible for their behavior. When the argument is deconstructed, it appears to have three components: (1) a principle that "it is a fundamental condition of fighting a just war that someone may be held responsible for the deaths of enemies killed in the course of it" (p. 67); (2) a prediction that humans will put into warfare autonomous military robots; and (3) a prediction about the autonomous nature of these future robots. When it comes to the autonomous nature of robots, Sparrow recognizes different conceptions of autonomy. He points out that the term is sometimes used to refer to weapon technologies already used in the battlefield (e.g. a cruise missile). Here autonomy seems to only mean independence from immediate human control. These weapons, he argues, are not different from other modern long-range weapons in terms of the ethical issues they raise. His concern, however, pertains to the next generation of intelligent robots, which he assumes to be autonomous in a different way. Based on claims made by some prominent Artificial Intelligence (AI) researchers, Sparrow assumes that future robots will make their own decisions in an 'intelligent fashion' and will therefore not be predictable. They will base their actions on their own internal 'desires', 'beliefs' and 'values' (p. 65). These features would make these technologies unpredictable and therefore assigning responsibility problematic. The unpredictability is not mere randomness, according to Sparrow, and the autonomy of the systems cannot be captured by "the mere fact that they are unpredictable" (p. 70). However, it remains somewhat "mysterious", even by Sparrow's own admission, what the autonomy of these systems entails and what distinguishes them from other 'dumb' and unpredictable weapons (p. 71).

Although Sparrow uses scare quotes in the descriptions of the next generation robots, he uncritically accepts the claims of the AI researchers. In doing this, Sparrow passes over the matter of what it means for a machine to have internal desires, beliefs, and values. He acknowledges that autonomy is a poorly understood and contested concept, but he resolves to assume that if robots as described by the AI researchers are developed, they will "have a strong claim to be autonomous as well" (*ibid*). Taking this underspecified conception of autonomy, he connects it to a moral philosophical conception. Sparrow writes: "To say of an agent that they are autonomous is to say that their actions originate in them and reflect their ends. Furthermore, in a fully autonomous agent, these ends are ends that they have themselves, in some sense, chosen" (*ibid*). Again he passes over what it means for a machine to have ends and to choose ends. He, thus, embraces a particular, yet black-boxed, conception of future autonomous robots, one that originates in an analogy with human autonomy in moral philosophy.

Sparrow seems to recognize that autonomous robots could be designed or used in ways that keep humans 'in the loop' but he is skeptical that this will happen. He predicts that humans will choose *not* to keep humans in the loop because the tempo of battle and the costs associated with keeping operators in the loop will pressure humans into choosing otherwise. He writes: "Weapons that require human oversight are likely to be at a substantial disadvantage in combat with systems that can do without. Thus, as soon as one nation is capable of deploying AWS [Autonomous weapon systems] that can operate without human oversight then all nations will have a powerful incentive to do so." (p. 69)

Sparrow might be right about this, but there is no way of knowing. His claim is a prediction and highly speculative. How humans will behave in the future depends on a whole host of factors that may or may not come into play as autonomous robots are developed and put into use. For example, Sparrow's paper might generate (or contribute to generating) a public conversation that convinces the public to protest the use of autonomous military robots; this, in turn, might convince political and military leaders not to put autonomous military robots into battle or to use them in limited ways. His paper might even be interpreted as an attempt to warn us not to develop and use such weapons. Of course, that military leaders and nation states might 'throw caution to the wind' and opt to use autonomous weapons without human oversight to kill enemies is one possible trajectory of the future, though it is only one of many. It is also possible that the human actors involved will choose to keep human oversight and to hold on to decision-making on the battlefield.

Generally, technologies are designed according to specifications and tested for reliability and predictability. Admittedly, knowing when something has been adequately tested is a complex matter and technologies are, indeed, sometimes released too soon. Nevertheless, this is precisely where relying on a vague, metaphorical account of machine autonomy may misdirect the discourse. How humans will make use of autonomous robots depends on how they operate and what sort of reliability and predictability can be achieved. Hence, if one wants to speculate about future trajectories, one has to think through the issues of reliability and whether or how autonomous robots will be amenable to reliability testing depends on how they operate. So, even if we stay in the realm of predictions, knowing what is likely to happen depends on what autonomy means in the case of robots. Are the operations deterministic or non-deterministic? What kinds of constraints on robot operations can and should be programmed in? What kind of limits will there be on reliability testing? Without attention being given to reliability, it is far from clear that the pressures of competitive warfare will lead humans to put robots that they cannot control into the battlefield without human oversight. And, if there is human oversight, there is human control and responsibility.

Sparrow's prediction that humans will put autonomous robots for which no human can be responsible (because they make their own choices and their actions originate in them) into battle is based both on his prediction that humans will decide to deploy them and also on his prediction that such robots might exist. His argument is hypothetical: *if* robots for which no human can be responsible become available, *then* humans will deploy them in warfare, and that will violate his just war principle. That the argument is hypothetical is supported by Sparrow's own reluctance to claim that existing and future robots are or will be fully autonomous. He writes that he will remain "agnostic on the question of the extent to which existing or future AWS can truly be said to be autonomous" (p. 66).

The significance of this hypothetical argument depends then on the plausibility or likelihood of robots for which no human can be responsible. The idea of such robots—that such robots might become available—is either a premise of Sparrow's argument or a prediction. If it is a premise, then the argument is hypothetical and circular. Since he treats autonomous robots as robots for which no human can be responsible, then if such robots were put into use, they would violate his just war principle. That is, by definition robots for which no human can be held responsible violates the principle that just war requires that someone may be held responsible for deaths. On the other hand, if robots for which no human can be held responsible is a prediction, then it is a prediction not about, or not just about, the development of robots, but about what humans will decide. Humans will decide whether robots for which no humans can be responsible ever come to be.

Whether such robots ever come to be depends both on how robots are designed and what sort of responsibility practices humans develop to make use of the robots. In developing autonomous robots and putting them into use, decisions will be made by humans as to how the robots will operate, when they are reliable enough to be put to what use, how they are deployed, with what level of oversight, etc. So, whether or not robots for which no human can be responsible become available will be the result of human choices, human choices both in the sense that humans will design the robots and the context of their use, and in the sense that humans will draw the conclusion as to whether humans can or cannot be held responsible for the robots. Yet, how and where these choices are made are questions that remain unaddressed and hidden from view in Sparrow's argument.

The argument that Sparrow makes in his Killer Robot paper illustrates what happens when metaphorical claims about machine autonomy are taken literally. He focuses his attention on autonomous robots as robots for which no human can be responsible as if that outcome were the only possible trajectory of the technology's development. The problem is that the plausibility and possibility of such entities is not a matter of logic or even technological evolution. Whether or not such robots come to be will be a human decision made in the design and use of robots. Unfortunately, attention to robots for which no human can be held responsible distracts attention away from the important human decisions that will shape the design and use of autonomous robots.

Robots will be responsible

Another position taken in the discourse on responsibility and autonomous robots is to consider that autonomous robots could in the future be held responsible for their own behavior. This position is rarely defended in depth; rather, the idea is entertained as a possibility arising from increased autonomy. For example, in a thorough analysis of the future of autonomous robots done for the Office of Naval Research, Lin et al. (2008) write: "we would be hard-pressed to assign blame today to our machines; yet as robots become more autonomous, a case could be made to treat robots as culpable legal agents." Legal responsibility is different from moral responsibility but, still, even to consider robots culpable or an entity of consideration in legal processes is a big step. It is a step from thinking that robot autonomy is like human autonomy (in certain respects) to thinking that robot autonomy is or could be equivalent to human autonomy for moral and/or legal purposes. The step is facilitated, at least in part, by keeping what goes on inside robots (that which is referred to as autonomy) as a black box.

Asaro acknowledges the contentious nature of the concept of autonomy and suggests a continuum of autonomy ending in 'full autonomy'. He suggests that in the future fully autonomous robots will not only be able to think, sense, and decide for themselves, but they might also "acquire moral capacities that imitate or replicate human moral capacities" and "be capable of formulating their own moral principles, duties, and reasons, and thus make their own moral choices in the fullest sense of moral autonomy" (2008, p. 2). In identifying a list of capacities, Asaro begins to dig deeper into the nature of the envisioned robots. However, each of the capacities is described in metaphorical terms. Robot thinking, sensing, and deciding are all terms used for human behavior that are being extended to robots even though what robots do—how they operate—is or is likely to be very different from what humans do.

There might be good reasons in the future for making robots legal entities of some kind. There might be good reasons for considering robots moral entities (Johnson 2006) or considering them agents in certain contexts (Grodzinsky et al. 2008). It makes sense to refer to certain kinds of behavior in robots as autonomy for particular purposes and in particular contexts. For example, when a robot vacuum cleaner cleans our floors, it makes sense (and is fun) to call the robot our housekeeper. Of course, we know that though they have a similar outcome with respect to our floors, a human housekeeper and our robot vacuum operate quite differently. When autonomy in robots is underspecified or left in a black box, it is easy to forget the differences between the way humans operate and the way robots operate.

Although the idea that current technological developments will eventually lead to human-like autonomy is a dominant narrative in the discourse on autonomous robots, the various levels and kinds of automation that we have seen in the last century have never resulted in machines that did exactly what human beings do. Automated systems take over part of a process previously done by human actors, performing it perhaps more efficiently, faster, or with greater power, though not in the same way as human actors.

Technologies do not merely replace human beings; rather they complement and change human activity (Parasuraman and Riley 1997). They allow human actors to do things they could not do before, and as a result they shift roles and responsibilities and create new ones. So it is with robots. Tracing the distinctive ways in which robotic systems perform tasks is essential to understanding how tasks and responsibilities are created and distributed across the broader sociotechnical system. For technologies that do not yet exist, however, we can only look at the negotiations currently taking place about machine autonomy and what autonomous robots should do in relation to human beings. Examining these negotiations helps to make explicit the choices that are currently being made about the distribution of tasks and that frame and shape the adoption of responsibility practices.

Negotiations around machine autonomy

Machine autonomy remains an elusive and ambiguous concept even in computer science and robotics. Researchers have diverging conceptions of machine autonomy as they use the concept for different purposes (Noorman 2009). In his book *Autonomous Robots*, Bekey (2005) takes autonomy to refer to "systems capable of operating in the real-world environment without any form of external control for extended periods of time"(p. 1). This rather general definition is intended to cover human and biological systems, as well as robotic systems. He maintains that most robots are not "fully autonomous", as they are not capable of surviving and performing useful tasks in the real world for extended periods of time. Yet, he does argue that some robots can be said to have autonomy at some lower level, as they are capable of operating in structured environments without any human intervention. Other computer scientist have offered more specific conceptions of machine autonomy, highlighting particular aspects by drawing analogies with capacities that make human beings or biological systems autonomous, such as independence from other agents in decision-making processes, the ability to generate goals, or the degree to which the agent can give rules to itself (Elio and Petrinjak 2005; Luck et al. 2003). Then there are those who argue that conventional notions of machine autonomy place too much emphasis on the computational system as an isolated entity; they propose definitions that highlight different aspects of human autonomy, such as the ability to understand the moral and social significance of certain actions (Murphy and Woods 2009; Falcone and Castelfranchi 2001).

The variety of conceptions of machine autonomy in the discourse on autonomous robots reflects the many ideas, ambitions and goals of the various social groups with a stake in the development of these technologies. Actors involved in the development of these technologies use their own conceptions to set specific goals to strive for or highlight particular aspects of their designs. For example, they might describe software agents as being autonomous in order to differentiate agent-based programming from other more conventional ways of programming (Luck et al. 2005). As a result of the many conceptions of machine autonomy, some would argue that autonomy has already been achieved, while others, including the U.S. Department of Defense (DoD), argue that machine autonomy is still a goal to strive for. The concept is, thus, still very much under negotiation.

Of all the social groups interested in the development of autonomous robots, the U.S. armed forces are highly influential in shaping the meaning of machine autonomy. Organizations within the armed forces, characterized by a strong emphasis on command and control, pressure for particular conceptions of autonomy that configure it as a bounded and measurable dimension of a technological system, rather than an unlimited ability to freely choose how to act. A closer look at these conceptions suggests that there are important questions to be asked about responsibility and autonomous robots other than the question whether or not human beings can still be held responsible for them.

The discourse of the U.S. armed forces suggests that the forces have embraced the idea of autonomous robots, and are pushing for even more autonomy. The Department of Defense's Research and Engineering Enterprise lists autonomous vehicles as one of five priorities. The Army, Navy, Air Force and Marine corps each have their own programs for developing more autonomous vehicles and since 2000, the Office of the Secretary of Defense has published six Roadmaps for the development of unmanned vehicles. These documents specify more autonomy as a key objective (see for example U.S. DoD 2009, 2011).[3]

Despite this strong commitment to the use of autonomous technology, there are diverging ideas about what the increased autonomy would entail. Some advocates suggest that autonomous robotics will 'take human actors out of the loop' completely. For example, (echoing Sparrow's concern) Adams (2001) claims that the growing need for agility, speed and information will push the human out of the loop. He argues that technologies will become so fast and generate so much information that human involvement will make these systems vulnerable. As he puts it, "the military systems (including weapons) now on the horizon will be too fast, too small, too numerous, and will create an environment too complex for humans to control" (p. 58). In other narratives, the autonomy of robots does not mean that human actors are out of the loop. Human actors may still be involved in decision-making processes that autonomous robots execute. As explained in the 2009 DoD Roadmap: "First and foremost, the level of autonomy should continue to progress from today's fairly high level of human control/intervention to a high level of autonomous tactical behavior that enables more timely and informed human oversight" (p. 27). Here machine autonomy seems to refer to robotic systems that operate in support of and in close communication with human actors. So, although the military community is invested in the idea that autonomous robots will have an increasingly important role in the future, what this will entail is still subject to negotiation.

In the negotiations about the meaning of autonomy, one dominant theme is to define the concept in terms of the degrees or levels to which a robotic system operates without direct human intervention. Such definitions are generally intended to make autonomy a measurable property of a robotic system (Elliott and Stewart 2011). Because it is a goal targeted to be reached in the near future, the military wants to measure progress towards that goal. This creates the need for a metric that evaluates, classifies and ranks new technologies in terms of their autonomous capabilities. In search of a metric, the Department of Defense Joint Program Office (JPO), the U.S. Army Maneuver Support Center, the National

[3] In its Report on Technological Horizons, the Office of the Chief Scientist of the U.S. Air Force concludes that the single greatest theme to emerge from the report "is the need, opportunity, and potential to dramatically advance technologies that can allow the Air Force to gain the capability increases, manpower efficiencies, and cost reductions available through far greater use of autonomous systems in essentially all aspects of Air Force operations" (2010, p. ix).

Institute of Standards and Technology (NIST), the Army Science Board and others have each offered their own set of levels of autonomy or robotic behavior (Huang 2008, Huang et al. 2003). In several Roadmaps, the DoD uses a scale of Autonomous Control Levels (ACL) that ranges from remotely guided to fully autonomous swarms. The various military services have developed similar scales with the aim of breaking autonomy down into identifiable pieces.

The different proposed descriptions of levels of autonomy all serve to make autonomy a measurable property of robotic systems. They typically define a taxonomy by listing in discrete steps the assumed capabilities that would allow a robotic system to perform a progressively larger part of a particular process independently. The descriptions of the levels are to be used to compare and evaluate new autonomous technologies. The ACL scale used by the DoD, for instance, is a simplified version of the chart developed by the Air Force Research Laboratory (AFLR) (Clough 2002). Researchers from the AFLR developed a chart based on the OODA (Observe, Orient, Decide and Act) loop, a conceptual device regularly used within the military to analyze decision-making cycles in terms of four sequential steps (Boyd 1987). For each step in the loop, the researchers defined a scale with increasing levels of autonomy. The levels of autonomy for the *Decide* step, for instance, describe the extent to which an unmanned aerial vehicle (UAV) is capable of making decisions based on the data and information to which it has access. This scale ranges from a UAV not making any decisions, to onboard trajectory planning and avoiding collisions, to tactical group planning and choosing tactical targets. The descriptions of the various levels on all four scales defined by the AFLR make machine autonomy a measurable functionality of a UAV.

One thing that the various proposed descriptions of levels of autonomy have in common is that they are relative to particular, often well-understood and clearly circumscribed, *processes* that the system is intended to perform. Having been developed for particular application domains and unmanned systems (e.g. ground, aerial), the scales reflect the ideas of their creators about the kinds of tasks and missions that the system is to perform and about the type of environments in which the system is supposed to operate (Huang et al. 2003). The scales of automation measure the extent to which a robotic system is capable of performing these tasks with limited or no guidance from a human operator.

In this way, they echo more conventional notions of machine autonomy as the high-end of an increasing scale of automation. That is, this conception of machine autonomy, as the measurable dimension of a technology that allows it to perform particular tasks unassisted by a human operator, is reminiscent of various taxonomies or levels of automation. Indeed one of the most prominent taxonomies, Sheridan and Verplank's (1978) scale of automation, inspired many of the proposed levels of autonomy that we currently find within military organizations.[4]

Sheridan and Verplank's seminal work illustrates how conceiving of machine autonomy as a measurable property of robotic systems can be compatible with human control. Sheridan and Verplank introduced their scale of automation to demonstrate the levels of control that can be shared between human operators and computers (Sheridan and Verplank 1978; Sheridan 1992). They consider automation to be the mechanization of well-defined processes, in which routine tasks are translated into some formalized structure that allows human operators to delegate some level of control to the automated system (Sheridan 1992). Automatic systems on the lower end of the scale leave decision-making and control to the human. Higher on the scale are systems that limit the choices a human actor can make in particular processes. For example, when the process to be automated is driving a car, an automatic gear shifter occupies a position somewhere on the lower end of this scale, as it only takes over the task of shifting the gear at the appropriate time. The driver still has to make decisions about the appropriate speed, when to break, and how to avoid obstacles. Higher levels of automation are attributed to those systems that close the control loop over the process. They are able to perform more tasks in the process and they further limit the decisions that the human operator makes. Thus, in a self-driving car equipped with navigation software, sensors and obstacle avoidance algorithms, the driver only has to monitor the behavior of the car, as it is capable of navigating its way through (certain kinds of) traffic at varying speeds. Sheridan and Verplank consider an automated system to be autonomous when it is left to perform all the steps in a particular process on their own, i.e. humans have neither the need nor the ability to intervene. Machine autonomy for them is, thus, about closing the control loop over a particular process. It is therefore bounded; it extends only as far as the process does.

Because Sheridan and Verplank's conception of machine autonomy is bounded, machine autonomy is compatible with human control. Human actors define the process that the machine is to perform and set constraints on what counts as acceptable machine behavior. When a machine's behavior transgresses these constraints, its behavior is considered a flaw or failure. Moreover, human actors specify the conditions under which automated

[4] A Task Force of the U.S. Defense Science Board defined autonomy as "a capability (or a set of capabilities) that enables a particular action of a system to be automatic or, within programmed boundaries, "self-governing."" (U.S. DoD 2012).

systems are allowed to perform tasks without human intervention. They draft protocols and regulations for appropriate use and create training programs to make sure that operators understand the possibilities and limitations of the system. However, while Sheridan's and Verplank's scale of automation is compatible with human control, it is—much like the autonomy levels defined by the various military services—blind to the work done by all these other human actors, because it only looks at the control relationship between human operators and the automated system.

Consider the US Navy's Phalanx system.[5] The Phalanx close-in weapon systems (CIWS)—designed to be the last line of defense against anti-ship missiles—is able to automatically search, detect and track missiles headed towards the ship and evaluate the threat. Once it has been turned on and it detects a target, it can, in principle, automatically fire. This system has a high ranking on Sheridan and Verplank's scale for the process of finding, searching and engaging targets. It closes the control loop over this process in the sense that the human operator has limited opportunity to intervene in how the Phalanx executes the process and cannot direct it to a different target once it is in operation. Yet, the development of this technology was the result of a careful crafting of the various components that make up the system. Software developers had to analyze and model the processes involved in locating, tracking, and engaging incoming missiles based on radar data, and they had to translate these processes into algorithms. In collaboration with users and military authorities they determined the appropriate parameter settings and acceptable thresholds that the algorithms require in order to make decisions about, for instance, whether or not a detected object is a missile heading towards the ship. To make sure the system would work as intended and within the limits of acceptable behavior, the software had to be verified, validated and subjected to numerous experiments and tests. Moreover, protocols and training programs had to be developed for the appropriate use of the system. And when accidents occurred, the technology and protocols had to be reevaluated and fine-tuned. So, although the system is able to perform independent of direct human control for a period of time, human actors defined the processes, set the constraints on appropriate behavior, and determined the conditions under which the system could be used. In this way, human actors are in control of how the system performs and what risks are acceptable.

Still, it would be misleading to suggest that all military actors in the discourse on autonomous robots think of autonomy as equivalent to high levels of automation, for some sharply distinguish machine autonomy from automation. For example, some argue that unlike automatic systems, autonomous robots (of the future) will only have to be instructed as to what to do, not how to do it (Marra and McNeil 2013). They assume that human operators and designers will not have to fully specify in advance the behavior sequences that a machine initiates in response to a particular input. In its 2011 Roadmap, the DoD argues that *automatic* systems are "fully preprogrammed and act repeatedly and independently of external influence or control" (2011, p. 43). They are able to follow a predefined path while compensating for small deviations caused by external disturbances. In 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" (*ibid*). Their behavior in response to certain events is not fully specified or preprogrammed. According to the DoD "[a]n autonomous system is able to make a decision based on a set of rules and/or limitations. It is able to determine what information is important in making a decision" (*ibid.*).

This conception of machine autonomy seems to imply that autonomous robotic systems would somehow be more flexible and unpredictable, as compared to automated systems, in deciding how to proceed, given predefined goals, rules or norms. Such a conception may give some the impression that human operators as well as developers would have less control over the behavior of the system. The machine not only operates independent of the human operator, but also, to a certain extent, independent of its human creators.

However, even here, machine autonomy does not mean that machines are free in the choices that they make; the conditions for deciding on how to proceed are carefully set by human actors. Human actors exert their influence in at least three ways. First, much like in the Phalanx case, developers and designers delimit the problem that the robotic system is intended to solve and thus set constraints on its behavior. One of the capabilities that the DoD suggests in its 2011 Roadmap as required for more autonomous operation is multisensory data fusion (MDF). MDF is necessary for processes like path planning, obstacle detection and tracking as well as map building. In order for a robot to better understand its surrounding, data from various sensors needs to be combined and converted into meaningful information. MDF algorithms tend be based on probabilistic methods or machine learning techniques. Such algorithms model the behavior of a particular physical system or environment based on assumptions about the characteristics and uncertainties of that system or domain (e.g. is it a linear dynamic system?) as well as the characteristics of the available sensors (Khaleghi et al. 2013). Such systems are able to operate more flexibly than a preprogrammed deterministic algorithm because they allow

[5] See http://www.raytheon.com/capabilities/products/phalanx/.

for variations and can respond to certain unforeseen contingencies. Yet, this flexibility is a function of the problem definitions that the developers or programmers of the algorithm have formulated. Although they do not have to specify all the possible situations the system may encounter, they generate a model that approximates the behavior of particular aspects of the world and their uncertainties based on prior knowledge and experience.

A second way that human actors exert their influence on autonomous robots that are somehow more than automatic systems, is through norms and rules; even in future systems, norms and rules will still govern the behavior of autonomous systems. The DoD, for example, in its 2011 Roadmap, expects machine autonomy to involve machines that adhere to laws and strategies provided by human actors. In the autonomous systems envisioned by DoD machine behavior could vary as long as it stays within these predefined constraints (2011). Note that this would be a remarkable feat, as it would require these robots to interpret laws and strategies and apply them appropriately in ever changing sociotechnical contexts.

A third way that human actors exert their influence on autonomous robots has to do with predictability. Conceiving of autonomous robotic systems as somehow more flexible and nondeterministic than conventional automation calls for an increased emphasis on reliability and trust in technology, and the need to develop better methods for verification and validation (V&V). The autonomous systems envisioned in the DoD's 2011 Roadmap would only be allowed to operate autonomously if they exhibit predictable and reliable behavior. A helicopter would be allowed to fly into an unknown environment avoiding obstacles and threats if the software controlling the helicopter would adhere to certain expectations and norms. For instance, it should not fly into trees, it should execute given instructions and it should fly between way points in a limited amount of time. The DoD states in this Roadmap:

> To ensure the safety and reliability of autonomous systems and to fully realize the benefits of these systems, new approaches to V&V are required. V&V is the process of checking that a product, service, or system meets specifications and that it fulfills its intended purpose (p. 50).

An emphasis on verification and validation can also be found in the Technological Horizons report of the Air Force: "Achieving these gains will depend on development of entirely new methods for enabling trust in autonomy through verification and validation (V&V) of the near-infinite state systems that result from high levels of adaptability and autonomy" (p. ix). The authors argue that although it is possible to develop systems with relatively high levels of autonomy, it is the lack of suitable V&V methods that stands in the way of certifying these technologies for use (U.S. Air Force 2010, p. IX). They claim that in the near- to mid-term future developing methods for "certifiable trust in autonomous systems is the single greatest technical barrier that must be overcome to obtain the capability advantages that are achievable by increasing use of autonomous systems" (p. 42). The assumed nondeterministic character of future autonomous systems, thus, creates a demand for new ways of predicting and understanding and for controlling these technologies. This kind of control of autonomous systems is partly in the hands of those who develop V&V methods or other methods of ensuring trust and confidence in these systems.

The need for reliability and predictability as well as the desire to constrain and regulate envisioned autonomous systems is not surprising given the hierarchical nature of military organizations. Within these organizations responsibilities are explicitly and formally distributed along a chain of command according to international and national laws and regulations. Such laws and regulations make clear who has the authority to make certain kinds of decisions and who should be held accountable for the outcome of these decisions. Military leaders and commanders are assigned responsibility even if they do not directly control the outcome, because they are accountable for setting and creating the conditions under which their subordinates act. Command responsibility, for instance, is a guiding principle in military organizations and operations, where commanding officers are required to sign off on decisions and assume responsibility for the units under their command. This means that those in charge set parameters on appropriate behavior and have a form of strict liability with regard to the actions of those under their command. Similarly, commanders and military leaders have a responsibility in formulating the rules of engagement, which includes specifying how and which weapons may be used, and they may be held to account for their decisions. The hierarchical distribution of responsibility constrains the use of unpredictable autonomous technologies. A commander in such an environment would be reluctant to allow his or her subordinates to deploy a robot that they know is unpredictable for fear that they would be held responsible for violating the laws of armed conflict as a result of the robot's rogue or unethical behavior (Schulzke 2012).

None of the various conceptions of machine autonomy described above imply that human actors are not in control of the technology they create and deploy. Rather, making robots autonomous in various ways means that human actors have different kinds of control. Human actors exert their influence as they choose the mathematical and probabilistic models that will guide the behavior of the robotic system; as they formulate restrictions on the conditions for use and specify and verify the levels of reliability and predictability that robotic systems need to exhibit. Designers, developers, human operators as well as managers, regulators and policy makers, thus, set constraints on what robotic systems can and cannot do. The question, then, is how will responsibility for these systems be distributed among the human actors who are essential to the development and operation of autonomous systems.

Negotiating responsibility strategies

The view (represented by Sparrow) that no human actors can be, or will in the future be, responsible for the behavior of autonomous robots results from use of a conception of machine autonomy that draws an analogy with human autonomy. In effect, the autonomy of robots is treated as a black box, leaving its workings opaque, and then drawing out implications as if machine autonomy were the same as human autonomy. The assumption is that the robot will do things on its own volition. However, when the black box is opened up and we see how autonomy is understood and 'made' by those involved in the design and development of robots, the responsibility questions change significantly. The important question is not whether human actors can be held responsible (they can), but how tasks are distributed among human and non-human components of the system, whether the machine parts have been adequately tested, whether the human actors involved have been adequately trained for their tasks, what risks are involved, and how those risks are being managed and minimized.

As shown in the previous section, opening up the black box reveals that machine autonomy is not a single idea. There are a variety of conceptions, and each has different implications for responsibility. Even when autonomous systems are understood to be capable of adapting to their environment and learning from experience, the uncertainty in the algorithms that govern the behavior of these systems does not necessarily mean that no human actors can be held responsible. 'Adaptive' and 'learning' are metaphors too and need to be unpacked. Probabilistic or machine-learning algorithms that enable a robotic system to adapt and learn do not just exhibit random behavior, they are designed to perform particular tasks and the extent to which their behavior can vary is constrained by their human developers and operators. Responsibility questions for these technologies have to do with the decisions and strategies employed in designing the system or in the operation of the system. Did the designers of the system construct accurate models of the problem domain? Did they provide an appropriate interface for human actors to interact with the system? Did they adequately test the system? Did those that deployed and used the system sufficiently take the known risks into account? These are questions that will come up when something goes wrong and we want to trace back who or what is at fault. They also come up when thinking about how to design these systems so that they are safe and can be accounted for.

Human actors can, thus, be assigned responsibility. Responsibility can be assigned to those who decide where and how robotic systems will be deployed and to those who validate and verify the behavior of the autonomous system. A human commander or operator still provides a goal for the robot to perform, and people have been involved in translating that goal into robot behavior. Developers of the software that control the robot define the possible range of behavior that robots can exhibit. How and when responsibility can be assigned to these people and what practices can be developed to do so should be part of the current discussions about the future of autonomous robots.

When it comes to complex human–machine ensembles, including autonomous robots, there are established practices to assign and ascribe responsibility (Noorman 2013). These *responsibility practices* are both backward- and forward-looking but they work together and inform each other. Forward-looking responsibility involves decisions about which tasks and duties are going to be assigned to which individuals or non-human components, that is, who will be responsible for what. Backward-looking responsibility involves tracing back where precisely an error or errors occurred after an untoward event has taken place. The fault may lie in how the software was programmed to behave, how human actors in various roles behaved, the comprehensibility (friendliness) of the interface between the human actors and hardware, and so on. Backward-looking responsibility generally relies on or at least presumes something about forward-looking responsibility. That is, when we understand how tasks and responsibilities were assigned in a system, it helps us to understand what went wrong. Also, when we trace back the cause of a failure we may discover that something else should have been, and in the future should be, delegated to a human or non-human component. For example, we may discover that another human should be put in a particular loop.

None of this is to say that the responsibility issues of autonomous robots are easy to identify or address. The complexity of autonomous robotic systems involves complex technological components, many human 'hands', and human–machine interfaces, and this means responsibility is distributed broadly. Thus it can be a daunting challenge to trace back who or what is at fault when something goes wrong (Johnson and Powers 2005). Investigations often lead in many directions since accidents can be caused by human error (bad judgment, negligence), failures in the artifacts, or failures in the human–machine interfaces. Moreover, for many automated systems no single human operator or developer can fully comprehend or directly control what the whole system does. In closed environments where conditions can be controlled, automated behavior may be predictable, but human–machine systems often operate in open environments where contingencies cannot be fully anticipated (Perrow 1999).

Our opening of the black box indicates that instead of asking whether human actors can be responsible for autonomous robot behavior, we should be focused on developing responsibility practices that work to minimize risk and that clearly establish lines of accountability. Such practices can incorporate strategies that are already part of established practices, including some of the strategies mentioned earlier such as strict liability law and validation and verification techniques. Responsibility practices also include assigning duties and obligations to human actors in the system and making sure that the human actors can manage the machines and in particular machine failure.

As military robots become more and more sophisticated and decision-making tasks are increasingly assigned to robots, it may be that new practices will have to be developed to ensure that the systems can be managed appropriately. That is, we have to negotiate about how to best assign forward- and backward-looking responsibility and what that entails. New strategies and mechanisms for holding human actors responsible will have to be negotiated. As more tasks are delegated to machines and responsibilities are distributed, human actors including those who test the behavior and set the constraints of acceptable behavior for the robots will likely have to acquire new skills and knowledge. Technologies that use probabilistic or machine-learning algorithms may exhibit more flexible behavior than conventionally automated systems. The people that deploy and use these technologies and interact with them will have to know how these technologies generally behave and what their possibilities and limitations are. Negotiations about these strategies are already underway in the development of autonomous robots.

Conclusion

Our analysis suggests that discourse about autonomy and responsibility is a nexus of negotiation about what autonomous robots will be. It shows that the development of autonomous military robots does not necessarily mean the end of human responsibility, as is sometimes suggested by narratives that claim that human actors will inevitably lose control over increasingly autonomous robots. There are currently many conceptions of machine autonomy that have different implications for the assignment and distribution of responsibility. Some of the proposed definitions found in military reports and Roadmaps describe machine autonomy as a measurable and bounded function or capacity of robotic systems. That is, autonomous systems can operate on their own for extended periods of time, but human actors are in control of how, when, and where they are allowed to operate.

This is not to say that we should stop being concerned about the tasks assigned to the non-human components of military robots. On the contrary, concerns about responsibility should continue to be part of the negotiations, and should shape the delegation of tasks to the human and non-human components of these systems. Instead of focusing on the question whether robots themselves or human actors can be held responsible for the behavior of robots, attention should be focused on the best allocation of tasks and control to human and non-human components and how to best distribute responsibility accordingly. The ascription of responsibility is therefore an integral part of the development and design of robots. Delegation of responsibility to human and non-human components is a sociotechnical design choice, not an inevitable outcome of technological development. Robots for which no human actor can be held responsible are poorly designed sociotechnical systems.

Acknowledgments This paper was written with support from the National Science Foundation under Grant No. 1058457. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank the anonymous reviewers for their helpful feedback on earlier versions of this paper.

References

Adams, T. (2001). Future warfare and the decline of human decision-making. *Parameters*, 31, 55–71.

Asaro, P. (2007). Robots and responsibility from a legal perspective. *Proceedings of the IEEE Conference on Robotics and Automation, Workshop on Roboethics*, April 14, 2007, Rome.

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.

Bekey, G. (2005). *Autonomous robots: From biological inspiration to implementation and control*. Cambridge, MA: MIT Press.

Bijker, W. E., Hughes, T. P., & Pinch, T. (1987). *The social construction of technological systems: New directions in the sociology and history of technology*. London, UK: The MIT Press.

Boyd, J. (1987). *A discourse on winning and losing*. Maxwell Air Force Base, AL: Air University Library Document No. M-U 43947.

Chopra, S., & White, L. W. (2011). *A legal theory for autonomous artificial agents*. Ann Arbor: The University of Michigan Press.

Clough, B. T. (2002). *Metrics, schmetrics: How the heck do you determine a UAV's autonomy anyway. Technical report*. Wright-Patterson AFB, OH: Air Force Research Lab.

Crnkovic, G. D., & Çürüklü, B. (2012). Robots—Ethical by design. *Ethics and Information Technology, 14*(1), 61–71.

Crnkovic, G. D., & Persson, D. (2008). Sharing moral responsibility with robots: A pragmatic approach. In P. K. Holst & P. Funk (Eds.), *Frontiers in Artificial Intelligence and Applications*. Amsterdam: IOS Press Books.

Elio, R., & Petrinjak, A. (2005). Normative Communication Models for Agent. *Autonomous Agents and Multi-Agent Systems, 11*(3), 273–305.

Elliott, L., & Stewart, B. (2011). *Automation and autonomy in unmanned aircraft systems. Introduction to Unmanned Aircraft Systems* (pp. 99–122). Boca Raton: CRC Press.

Falcone, R., & Castelfranchi, C. (2001). The human in the loop of a delegated agent: The theory of adjustable social autonomy. *IEEE Transactions on Systems, Man and Cybernetics, 31*(5), 406–418.

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2008). The ethics of designing artificial agents. *Ethics and Information Technology, 10*, 115–121.

Hellstrom, T. (2012). *On the moral responsibility of military robots*. Ethics and Information Technology (forthcoming).

Huang, H. (2008). *Autonomy levels for unmanned systems (ALFUS) framework volume I: Terminology version 2.0*. NISTSP 1011-I-2.0, National Institute of Standards and Technology, Gaithersburg, MD, September 2004.

Huang, H., Messina, E., & Albus, J. (2003). Autonomy level specification for intelligent autonomous vehicles: Interim progress report. In *Proceedings of the performance metrics for intelligent systems (PerMIS) workshop*, September 16–18, 2003, Gaithersburg, MD.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*(4), 195–204.

Johnson, D. G., & Powers, T. M. (2005). Computer systems and responsibility: A normative look at technological complexity. *Ethics and Information Technology, 7*(2), 99–107.

Khaleghi, B., Khamis, A., Fakhreddine, O. K., & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion, 14*(1), 28–44.

Lee, N., & Brown, S. (1994). Otherness and the actor network. *American Behavioral Scientists, 37*(6), 772–790.

Lin, P., Bekey, G., & Abney, K. (2008). Autonomous military robots: Risk, ethics, and design. http://ethics.calpoly.edu/ONR_report.pdf. Accessed October 14, 2011.

Luck, M., McBurney, P., Shehory, O., & Willmot, S. (2005). Agent technology: A roadmap for agent based computing (A Roadmap for Agent Based Computing), AgentLink, 2005. http://www.agentlink.org/roadmap/. Accessed February 12, 2014.

Luck, M., Munroe, S., & d'Inverno, M. (2003). Autonomy: Variable and generative. In H. Hexmoor, C. Castelfranchi, & R. Falcone (Eds.), *Agent Autonomy* (pp. 9–22). Dordrecht: Kluwer.

Marino, D., & Tamburrini, G. (2006). Learning robots and human responsibility. *International Review of Information Ethics, 6*, 46–51.

Marra, W. C., & McNeil, S. K. (2013). Understanding 'The Loop': Regulating the next generation of war machines (May 1, 2012). *Harvard Journal of Law and Public Policy,* 36(3). http://ssrn.com/abstract=2043131.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

Murphy, R. R., & Woods, D. D. (2009). Beyond Asimov: The three laws of responsible robotics. *IEEE Intelligent Systems, 24*(4), 14–20.

Nagenborg, M., Capurro, R., Weber, J., & Pingel, C. (2008). Ethical regulations on robotics in Europe. *AI & SOCIETY, 22*, 349–366.

Noorman, M. (2009). *Mind the gap a critique of human/technology analogies in artificial agent discourse*. Maastricht, The Netherlands: Universitaire Pers Maastricht.

Noorman, M. (2013). Responsibility practices and unmanned military technologies. *Science and Engineering ethics*. doi:10.1007/s11948-013-9484-x.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors Society, 39*(2), 230–253 (224).

Perrow, C. B. (1999). *Normal accidents: Living with high-risk technologies*. 2nd Edition, Princeton, NJ: Princeton University Press.

Schulzke, M. (2012). Autonomous weapons and distributed responsibility. *Philosophy and Technology*. http://link.springer.com/article/10.1007%2Fs13347-012-0089-0. Accessed December 14, 2012.

Sheridan, T. B. (1992). *Telerobotics, automation, and human supervisory control*. Cambridge, MA: MIT Press.

Sheridan, T. B., & Verplank, W. (1978). *Human and computer control of undersea teleoperators*. Cambridge, MA: Man–Machine Systems Laboratory, Department of Mechanical Engineering, MIT.

Singer, P. (2009). *Wired for war: The robotics revolution and conflict in the 21st century*. New York, NY: Penguin.

Sparrow, R. (2007). Killer robots. *Journal of applied philosophy, 24*(1), 62–77.

U.S. Air Force Chief Scientist. (2010). *Report on technological horizons: A vision for air force science & technology during 2010–2030*. Vol 1. AF/ST-TR-10-01-PR, May 15, 2010.

U.S. Department of Defense. (2009). *FY2009-2034 Unmanned systems integrated roadmap*. http://www.acq.osd.mil/psa/docs/UMSIntegratedRoadmap-2009.pdf. Visit September 20, 2011.

U.S. Department of Defense. (2011). *FY2011-2036 Unmanned systems integrated roadmap*. http://www.acq.osd.mil/sts/docs/UnmannedSystemsIntegrated-RoadmapFY2011-2036.pdf. Accessed January 3, 2012.

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 November 5, 2012.

Wallach, W. (2013). *Terminating the terminator: What to do about autonomous weapons*. http://ieet.org/index.php/IEET/more/wallach20130129 posted January 28, 2013; Accessed February 2, 2013.

Wallach, W. & Allen, C. (2013). Framing robot arms control. *Ethics and Information Technology*, *15*(2), 125–135.