Ethics Inf Technol (2013) 15:99–107
DOI 10.1007/s10676-012-9301-2

ORIGINAL PAPER

On the moral responsibility of military robots

Thomas Hellström

Published online: 9 September 2012

Abstract This article discusses mechanisms and principles for assignment of moral responsibility to intelligent robots, with special focus on military robots. We introduce the concept *autonomous power* as a new concept, and use it to identify the type of robots that call for moral considerations. It is furthermore argued that autonomous power, and in particular the ability to learn, is decisive for assignment of moral responsibility to robots. As technological development will lead to robots with increasing autonomous power, we should be prepared for a future when people blame robots for their actions. It is important to, already today, investigate the mechanisms that control human behavior in this respect. The results may be used when designing future military robots, to control unwanted tendencies to assign responsibility to the robots. Independent of the responsibility issue, the moral quality of robots' behavior should be seen as one of many performance measures by which we evaluate robots. How to design ethics based control systems should be carefully investigated already now. From a consequentialist view, it would indeed be highly immoral to develop robots capable of performing acts involving life and death, without including some kind of moral framework.

Keywords Moral responsibility · Robots · Military robots · Autonomy · Robot ethics

T. Hellström (✉)
Department of Computing Science, Umeå University, Umeå, Sweden
e-mail: thomash@cs.umu.se

Introduction

This article discusses possible mechanisms and principles for assignment of moral responsibility to intelligent robots. Special focus is on military robots, which are being massively introduced by a large number of armies around the world.[1] The military robots will become more and more autonomous and lethal, and responsibility issues must be discussed as well as extended to match the new reality. The article is organized as follows. Section 2 gives an overview of existing battlefield robots. In Sect. 3, these robots are classified based on *autonomous power,* a new concept introduced as an extension of the *autonomy* concept. This helps to identify the kind of robots that should come into question when discussing responsibility. Section 4 analyzes in what way *moral responsibility* may be applicable to robots, and how it relates to autonomous power. Assignments of moral responsibility in hypothetical war scenarios with and without robots are analyzed in Sect. 5, followed by a final discussion and conclusions in Sect. 6.

Robots in warfare

The last decade has seen an intense research and development of military Unmanned Ground Vehicles (UGVs) and Unmanned Aerial Vehicles (UAVs). The results are already put in extensive use in armed conflicts around the world (Wezeman 2007 p. 5–6, Singer 2009a, b). However, the use of military robots is not at all a new idea. For instance, UAVs have been used by several armed forces

[1] Between 50 and 80 countries either already use or are in the process of acquiring the technology to start using military robots (ABIresearch 2011).

since the 1960s (Wezeman 2007). The early UAVs were not armed but mainly carried out reconnaissance missions, and this is still the case for the majority of all UAVs. Models such as the MQ-9 Reaper are about to change this. A representative of the U.S. Air Force explains (U.S. Air Force 2006): "We've moved from using UAVs primarily in intelligence, surveillance and reconnaissance roles before Operation Iraqi Freedom, to a true hunter-killer role with the Reaper". The first large scale use of armed UAVs in combat operations was the attacks carried out by the U.S. against Taliban and Al Qaeda leaders in Yemen and Afghanistan in 2002 and 2003 (Bone and Bolkcom 2003). Still, all UAVs are mainly tele-operated, and the operator controls firing of arms. Future development plans for UAVs include multi-robot systems, either centralized (mothership style) or decentralized (swarm style). A mothership plane would launch a large number of UAVs, coordinate and control their actions and later recover them in the air. The swarm approach uses a large number of independent UAVs. Rather than being centrally controlled, each unit decides what to do on its own, resulting in highly effective overall results (Singer 2009c).

Several types of military UGVs are also extensively used. By the end of 2008, there were about 12,000 robots operating on the ground in Iraq (Singer 2009a, b). Most of these vehicles/robots are not armed but are used for sniper detection, disrupting or exploding explosive devices and surveillance in dangerous areas. One of the most common models for this kind of military operations is the PackBot (Yamauchi 2004) from iRobot (iRobot 2012), the manufacturer of the Roomba robot vacuum cleaner. This robot is equipped with cameras and communication equipment and an optional arm with manipulator. Its small size enables it to be carried by a soldier in a backpack. Through the operator's control unit, the soldier tele-operates the robot and views the output from the cameras and other sensors. The PackBot is typically used to enter buildings, report on possible occupants, and trigger booby traps. It may also be equipped with special sensors for sniper detection. The Bloodhound (Yamauchi et al. 2002), a new version of the PackBot, will be able to search for injured soldiers and provide some rudimentary treatment. Further development of this robot will include capabilities to drag an injured soldier to safety. The SWORDS (GlobalSecurity 2012) (Special Weapons Observation Remote Direct-Action System) robot made by Foster–Miller can be equipped with machine guns, grenade launchers, or anti-tank rocket launchers as well as cameras and other sensors. The robot is able to navigate on its own with a satellite navigation system (GPS). At present, a soldier navigates and fires onboard weapons by teleoperation from a safe distance away. While SWORDS robots have been deployed in Iraq, the scarce information available indicates that they have so far not fired a single shot in field (Sofge 2009). The manufacturer is currently working on a successor, the MAARS (Modular Advanced Armed Robotic System) robot (QineticQ 2012).

The Phalanx is an anti-ship missile system that has been in service since the 80′, primarily by the U.S. Navy (U.S. Navy 2011). It consists of a machine gun mounted on a rotating base. In automatic mode it searches, detects, evaluates, tracks, and engages without intervention of human operators. The decision to fire is taken by a computerized radar system estimating speed and direction for approaching objects. The Phalanx is being constantly upgraded and has reportedly been deployed in the navies of 24 nations (Raytheon 2009).

The SGR-1 Security Guard Robot from Samsung (Samsung 2012) is a stationary system developed for the Korean Demilitarized Zone. It is able to detect and identify targets in daylight and at night, using a combination of laser range finders, low-light high-resolution cameras, thermal cameras and infrared sensors. The system may be equipped with machine guns and an acoustic device that "emits a tone powerful enough to make intruders nauseous and drop to the ground" (Hildebrand 2009). This provides for either lethal or non-lethal response. The system can be operated manually but also has an automatic mode in which it fires on its own.

To summarize, the robots used in war zones today vary considerably regarding lethality and degrees of self-management. The vast majority of all military robots are tele-operated and not equipped with any kind of weapons. They are used for supporting functions such as surveillance, sniper detection and neutralizing explosive devices. Some robots, notably UAVs, are equipped with weapons controlled by tele-operation. A few ground based robotic systems that automatically fire are in use, primarily in restricted border areas and at sea. A few armed autonomous mobile robots have been developed and introduced to the battlefields, primarily for evaluation.

Classification of robots

In this section, an attempt to classify robots based on their degree of self-management and lethality is made. Self-management relates to the notion of *autonomy*, a word with many proposed meanings and definitions. During a research meeting organized by EUCogII in Groningen 2011 (http://www.eucognition.org), more than one hundred researchers from a wide spectrum of disciplines spent considerable time discussing their views on the matter. While no conclusive definition was neither anticipated nor reached, several keywords reappeared during the discussions: independence, self-ruling, internal agenda, internal

goal, and intention. Franklin and Graesser (1997) review a number of definitions and interpretations of autonomous agents that are used within the area of artificial intelligence, and end up with the following, often referenced, definition.

> An *autonomous agent* is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future.

Franklin and Graesser themselves notice that a thermostat satisfies all the requirements of the definition of an autonomous agent. Likewise, a landmine should be regarded as fully autonomous. Its triggering mechanism senses the environment and acts, according to its own agenda, by blowing itself to pieces in a way that certainly effects what it will sense in the future. The use of anthropomorphic expressions like "own agenda" clearly invites to different interpretations of the definition. It could be argued that the landmine is not at all very autonomous since it explodes not as a result of its own agenda, but rather the constructor's. However, it is hard to define what the expression own agenda really means without using equally anthropomorphic concepts like free will and determinism. In any case, the above definition and many other definitions of autonomy focus on independent self-ruling, but this ability alone is of limited value unless the agent also has the ability to perceive and act in its environment. For the upcoming ethical discussion, there is a need to identify and characterize autonomous agents with varying repertoires of physical actions, interactions and decision mechanisms. We therefore introduce the extended concept *autonomous power* as follows:

> By *autonomous power* we denote the amount and level of actions, interactions and decisions an agent is capable of performing on its own.

In this definition, autonomy is captured by the expression "on its own". While being purposely vague, it covers both lexicographic definitions and a common intuitive meaning of the word autonomy: the ability to act independently. The word "actions" refers to the ability to affect the world. The word "interactions" refers to the awareness of and adaption to what is going on in the world. Finally, the word "decisions" refers to the algorithm that controls how the agent acts in different situations. The "amount and level" can mean very many different things. For the decision part, it directly refers to the complexity of the control algorithm mapping sensor information to actions, and indirectly to qualities such as the ability to learn from experience. For a landmine, the action repertoire is limited to one simple act, namely exploding. The interaction with the environment is also simple and limited to sensing activation of a binary trigger mechanism. Finally, the decision process is limited and simple: the explosion is executed if and only if the trigger mechanism is activated. Hence, a landmine has low autonomous power, although it is fully autonomous.

Autonomy is related to degrees of automation as discussed for collaborative human-robot systems (Sheridan 1992). Parasuraman et al. (2000) propose four classes of functions in such a system: 1) information acquisition; 2) information analysis; 3) decision and action selection; and 4) action implementation. Depending on to what extent a robot performs these functions without human assistance, the human-robot system has a certain degree of automation. Just like most definitions of autonomy, the focus is on self-ruling, without directly taking into account the complexity of the task. For instance, information analysis may be very straightforward and simply, or it may require demanding situational awareness, experience, and learning. While running at the same degree of automation, robots with vastly different autonomous power are required for the two cases. This means that robot involvement for the four proposed classes of functions requires an autonomous power that depends not only on the chosen level of automation but also on the specific task. Sheridan (1992, p.356–357) incorporates complexity of the task (task entropy) in his thorough analysis of automation, but not explicitly in his scale of degrees of automation.

Our definition of autonomous power differs from regular autonomy by covering not only the ability to self-ruling but also the ability to perceive, analyze, and act. It is a measure with a range from low to high. At the upper end of the scale, an agent with high autonomous power affects the environment with a large variety of actions, controlled by an advanced internal decision process, based on extensive interaction with the environment. At the lower end of the scale we find agents and artifacts that might be very independent, but also do not interact very much.

In Fig. 1, a number of weapons and robots are classified with respect to *lethality* and autonomous power. By *lethality* we denote a weapon's or a robot's physical ability to kill, if being used or activated in the intended fashion.[2] It may, for instance, be thought of as the statistical expectation value of the number of casualties. Traditional weapons are gathered to the left in Fig. 1, indicating that the autonomous power is very low. However, already a crossbow, that autonomously transports the arrow towards the target, can be said to possess non-zero autonomous power. Regular bombs have a bit more autonomous power in the sense that they explode as a result of interaction with the environment, and are certainly more lethal than

[2] For instance, a regular aircraft like a Boing 767 is not considered a highly lethal weapon, although it may be used to kill thousands of people. Likewise, a kilogram of water, with an equivalent mass-energy corresponding to one thousand Nagasaki bombs, has low lethality when used in the intended fashion.

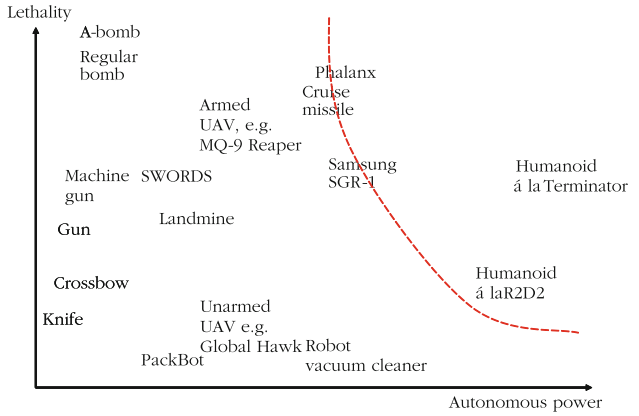

Fig. 1 Lethality and autonomous power for a selection of weapons and robots. Autonomous power denotes the amount and level of actions, interactions and decisions the considered artifact is capable of performing on its own. Robots to the right of the *dotted line* have the ability to kill as a result of complex internal decision processes

crossbows. However, in general, traditional weapons have a very low autonomous power compared to the new generation of military robots. The UAV Global Hawk can navigate autonomously, but has no weapons and hence low lethality. A cruise missile is an example of an old weapon that is actually both highly lethal and possesses a relatively high autonomous power compared to most other weapons. A Tomahawk cruise missile navigates by a combination of inertial navigation and matching of radar altitude profiles of the terrain with stored map references. Close to the target, navigation is done by matching stored camera images of the target area with actual camera images (Riedel et al. 2010). Systems like Phalanx and SGR-1 have still higher autonomous power since they detect, classify, and possibly fire as a result of fairly complex algorithms involving analysis of data from the environment. The still not (2012) realized humanoid robots á la Terminator and R2D2 have autonomous power approaching that of a human. However, even Terminator has lower lethality than for instance a Phalanx.

Note that the used concepts lethality and autonomous power are vaguely defined concepts. The values used to fill in Fig. 1 are based on an intuitive understanding of the definitions, and the axes are not necessarily linear. Thus, the exact locations of weapons and robots can certainly be discussed and the diagram should primarily be seen as an illustration of the general idea. The dotted line indicates an approximate boundary for robots that have physical ability to kill and enough autonomous power to do so as a result of internal autonomous decision algorithms working in interaction with the environment. Robots to the right of the boundary are the main focus of interest for the ethical discussion in this article. This does not mean that tele-operated robots, like the SWORDS robot (low autonomous power and medium high lethality), have not created new problematic moral issues. Neither does it mean that future versions of the Roomba robot vacuum cleaner (high autonomous power and low lethality) will be uninteresting from a moral point of view. Already existing unarmed UAVs used for surveillance raise serious ethical questions regarding privacy and integrity. However, all these kinds of robots fall out of the main scope of this paper since we are focusing on armed military robots. The intense ongoing research and development will no doubt introduce new robots to the right of the boundary, thereby further emphasizing the need for ethical considerations.

Moral responsibility

Returning to the main topic of this paper, we will in this section discuss the concept moral responsibility, how it relates to autonomous power, and in what way it can be applied to robots. Aristotle (384–323 BCE) was one of the first to construct a theory of moral responsibility. In *Nicomachean Ethics* III.1–5 (Aristotle 1985), an agent is described as morally responsible for an action, if it is worthy of praise or blame for having performed the action. Only a certain kind of agents can be ascribed of responsibility, namely those who possess a capacity for decision. Furthermore, the action has to be voluntary. According to Aristotle, a voluntary action has two distinctive features. First, the action must have its origin in the agent. Second, the agent must be aware of what it is doing. Along such lines, moral responsibility has been examined and discussed over the millennia, branching off into hard and open issues involving intention, consciousness, free will and determinism (for an overview see for instance Eshleman (2009)). Until recently, the discussion has mainly focused on living agents. Incorporating robots into the discussion certainly does not simplify the picture.

Some researchers view moral responsibility as governed by pragmatic norms of a group (Dennett 1973; Strawson 1974), and hold that moral responsibility should be seen as a social regulatory mechanism aiming at supporting actions considered to be good, and simultaneously suppressing actions considered to be bad. A classical view describes moral responsibility as consisting of two parts: causal responsibility and intention (see e.g. (Dodig-Crnkovic and Persson 2008)). It is sometimes claimed that advanced mental states, such as intention, are and will always be missing in machines, and that a robot for that reason could never be ascribed moral responsibility (Johnson 2006). However, this is by no means the only position among scholars, see for instance (Dennett 1997) and (Bechtel 1985). Asaro (2006) reasons that moral responsibility is a continuum from no to full responsibility, and points at how

this way of thinking already exists in society, for instance in the way we treat children as not always fully responsible for all their actions. For the continuing discussion these views are adopted. Hence, we do not engage in the more philosophical question of the objective nature of moral responsibility. Instead, moral responsibility will be regarded as a quality we assign to others, in varying degrees based on some, possibly unspecified, norm system. This view blurs the borders between moral responsibility and task, role, and legal responsibility. All these different types are related and somewhat overlapping (Stahl 2004, p. 105f), and legal responsibility is for instance affected by moral views in a society. While the clear focus of this paper is on moral responsibility, the discussion sometimes will touch on the other aspects of responsibility. The common and important thread is that responsibility is a quality we assign to humans, and possibly also to robots.

Already now, humans assign responsibility to non-living entities. Companies are *juristic persons* and can own property and sign deals with other companies or with natural persons. Furthermore, they can be sued, punished and even accused of immoral behavior, and are in many respects regarded as entities separated from both their employees and owners (who may be natural or juristic persons).[3] Several studies show how humans attribute responsibility to computers (Friedman 1990; Friedman and Millett 1995, 1997; Moon and Nass 1998). For instance, in a study by Friedman and Millett (1995), 21 % of the participants held a computer morally responsible for incorrect decisions in two scenarios; medical radiation treatments in which some patients were over-radiated, and evaluation of job seekers, in which the computer rejected qualified applicants. Other studies show how humans also have a tendency to attribute responsibility to mobile robots. Kim and Hinds (2006) conducted experiments in which people were asked if a mobile robot performing a delivery task was responsible for errors that were made, and if the robot was to blame for problems that were encountered in accomplishing the task. The results showed that robots were seen as responsible and worthy of blame, in proportion to their degree of autonomy that was varied between two levels. In the low-level mode, the robot asked the human for confirmation in one phase of the operation, and in the high-level mode, the robot decided by itself whether to continue or not. These levels correspond to degrees of automation (Sheridan 1992, pp. 356–357) and also to common definitions of autonomy or self-ruling. Just as we suggest autonomous power as an extension to the autonomy concept, we believe that autonomous power is more decisive than autonomy when assigning moral responsibility. Furthermore, autonomous power catches the key factors for assignment of moral responsibility to humans reported in the beginning of this section: the extent to which the human can affect the environment ("causal responsibility"), and the extent to which this influence is a result of an internal decision process ("possess a capacity for decision" or "intention"[4]). Transferred to the robotics domain, this corresponds to the concept of autonomous power, and leads to the following proposition:

> Our tendency to assign *moral responsibility* to a robot increases with its degree of autonomous power.

With reference to the classification of robots in Fig. 1, robots to the right of the dotted line are both potentially lethal and may to some extent be viewed as morally responsible for their actions. For robots to the left of this line it certainly does not make sense to talk about responsibility at all. For instance, the moral responsibility of a cruise missile is most likely regarded as zero by most people. The situation might be different if the missile would talk to the military commander while cruising, and for instance suggest changed target coordinates due to its observations on ground.

It is known that our willingness to assign moral responsibility to robots depends/will depend on several other factors than autonomous power, for instance how human-like the robots are. In experiments by Hinds et al. (2004), humans acting together with robotic co-workers retain more responsibility for their own work when the robots are less humanoid. Experiments by Kim and Hinds (2006) indicate that humans blame robots less if the robots are transparent, i.e. somehow explain their behavior to the humans. The final answer to how and why we assign moral responsibility to robots will have to wait until we have much more autonomously powerful robots than we have today. However, it would be possible to already today further investigate the mechanisms that control human behavior in this respect. Results from such investigations may be used when designing future robots, to control unwanted tendencies to assign responsibility.

In the following section, hypothetical war scenarios, in which responsibility is assigned not only to humans but also to robots, are described and related to the discussed issues. The purpose is to investigate the process of assignment of moral responsibility to humans and robots, and also to further support the proposed connection to the concept autonomous power.

[3] However, this can also be seen as an example of *collective responsibility* of the company's employees and owners. See for instance (Young 2010, p. 67).

[4] Note that equating the internal decision process with the concept of intention, very much is an observers perspective, consistent with the previously adopted view of moral responsibility as a quality an observer assigns to an agent.

Assigning moral responsibility in war

A simplified chain of command, and assignment of responsibility, for military actions is illustrated in Fig. 2. In a democracy, the People elect politicians to act on their behalf. At least at the top, this involves extensive assignment of responsibility. The politicians are given a large repertoire of powerful tools and large freedom to use them. For various unfortunate reasons, these politicians may decide on going to war. Overall tasks, and responsibilities, are then assigned to a large number of military commanders who in turn give orders, and responsibility, to lower ranked commanders and eventually to soldiers who execute the physical actions of war in field. The amount of responsibility is not the same for all agents: "The higher their rank, the greater the reach of their command, the larger their responsibilities." (Walzer 2006, p.316). The intimate connection between power and responsibility is well established also in politics in general (Connolly 1974). Overall, we find support for our thesis that the amount of assigned responsibility is related to the level of autonomous power of the receiving part: "the amount and level of actions, interactions and decisions an agent is capable of performing on its own". A politician often has a large repertoire of actions, and is also given a lot of freedom to act within a wide political frame. This corresponds to a high autonomous power, and indeed the politician bares a large responsibility for the consequences of his or her actions. A high military commander may be given a military task that has to be performed within certain given constraints. As a consequence, also the assigned responsibility is somewhat constrained. The commander may equip soldiers and give strict orders with very little room for interpretations and autonomous decisions. Hence, the commander assigns a relatively small amount of responsibility to the soldiers. In one scenario, the robots, created by scientists and engineers, are essentially tele-operated and can be seen simply as advanced weapons with low autonomous power. Hence, no responsibility is assigned to these robots.

In a world with autonomously powerful battlefield robots, scientists and engineers get orders and financial incentives to further develop and build advanced military robots. The robots may be equipped with lethal capabilities and be programmed to execute orders issued by their users. However, the orders may be formulated at a high and vague level, leaving a number of complex decisions and interpretations to the robot. Already today, a few systems such as Phalanx and the Korean SGR-1, are approaching such autonomous power (see Fig. 1). They are capable of detecting, identifying and firing as a result of an internal decision processes, albeit much more primitive than their human counterparts. Much higher levels of both autonomy and power are probably needed before such robots will be regarded as even partly responsible for their actions. If and when this happens, the chain of assignment of responsibility has to be extended in several ways to reflect this fact. Soldiers will start giving high-level orders, such as "Cover my back" or "Secure that hill", to the robots. With increased autonomous power, the military commanders may start treating the armed intelligent robots as human soldiers. The robots will receive orders and *rules of engagement* and will be expected to follow them in the same way as human soldiers do. The robots will be equipped with highly advanced cognitive abilities for perception, planning and learning, and also with a repertoire of complex behaviors and capabilities to operate lethal weapons of various kinds. In Fig. 2, dashed arrows illustrate this partial assignment of responsibility.

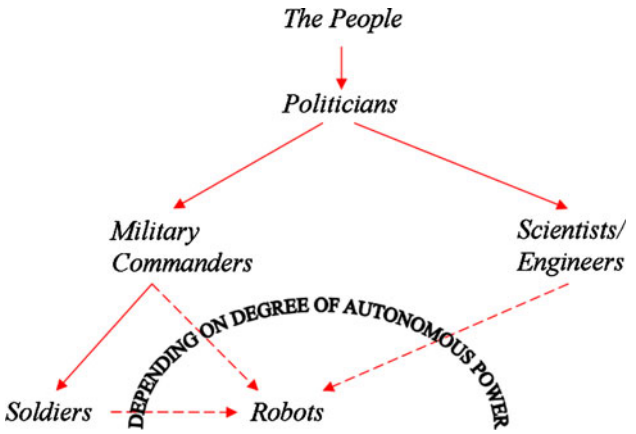

Fig. 2 Partial assignment of moral responsibility to one moral agent by another. With today's mainly tele-operated robots, no responsibility is assigned to the robots. This situation may change as autonomously powerful battlefield robots are introduced

With the tele-operated battlefield robots of today, responsibility is not assigned to the robots at all. Rather, the robots are much like regular weapons and one may even apply the slogan of the National Rifle Association (NRA): "guns don't kill people, people kill people". A tele-operated UAV is not lethal until a human operator presses the Fire button. However, the slogan is not directly applicable to robots such as Phalanx or the SGR-1 which may "decide by itself" (in the sense that no operator needs to press the Fire button) if and when to fire. To be applicable in a world with lethal autonomously powerful robots, the NRA slogan will have to be modified to something like "guns don't kill people, robots kill people".

Note that assignment of responsibility not necessarily is a zero–sum game in which a superior, who gives an order and assigns responsibility to a subordinate, is released from responsibility for the consequences of executing the order. Nor is a subordinate necessarily released from responsibility for his or her own actions just because he or she

follows order. Both ways of reasoning are examples of *diffusion of responsibility*. The first case is dealt with in the doctrine of *command responsibility* (Walzer 2006, p.316) that emphasizes a military superior's responsibility. The second case is a more open issue, where subordinates sometimes are released from responsibility. One extreme example is child soldiers who commit war crimes. It is often argued that they rather are victims than perpetrators, partly because of their "unique psychological and moral development" (Grossman 2007). In general, we often treat children as not fully responsible for their actions and in the future we may very well view robots, who indeed also have a unique psychological and moral development, in a similar way. Just as parents share moral responsibility with small children, military commanders and soldiers, and possibly also politicians, scientists/engineers and the People, may share moral responsibility for actions conducted by military robots.

Discussion and conclusions

The introduced concept of autonomous power, defined as a combination of self-ruling and capacity for actions, interactions and decisions, extends the traditional concept of autonomy such that weapons and military robots can be classified in a meaningful way for ethical considerations. Furthermore, autonomous power seems to be a decisive factor when assigning moral responsibility to other agents.

Already now, humans have a tendency to assign responsibility to computers and robots, and there are reasons to believe that this tendency will be strengthened in the future since the technological development leads to increasingly more autonomously powerful robots. Development of more and more advanced military robots may very well create situations in which it becomes natural to talk about moral responsibility of the robots, shared with or separated from their creators, military commanders and soldiers who put them to work. If, and to what extent, this will happen is indeed hard to know, and has to do with general development of artificial intelligence and robotics. However, we do not believe that questions concerning morally responsible robots have to, or should, wait until robots with near human-level intelligence are developed. Rather, the need to discuss these questions already exists and will for certain increase in the future.

The fact that we may be inclined to assign responsibility does not mean that we are justified in so doing, and there are definitely good reasons to affect and regulate these mechanisms as we incorporate advanced computers and robots in our society. Many researchers are worried about a future where robots are blamed as a way for humans to escape responsibility. An attempt to highlight the seriousness of the problem is taken by a group of researchers who developed and signed a set of rules emphasizing the responsibility of the individual researcher, engineer and user (Miller 2011).[5] This kind of worry is certainly motivated and should be taken seriously. It may indeed be problematic to blame an individual programmer in a team of maybe hundreds of people who developed the maybe millions of lines of program code that constitute a robot's brain. Furthermore, in the not so distant future, robots will, like humans, learn from experience such that their behavior partly depends on events out of control of both developers and users. Matthias (2004) argues that responsibility issues for advanced learning robots have to be seriously discussed both from a moral and legislation point of view. A further complication is *emergent behaviors* that unexpectedly may appear out of the complexity of several interacting behaviors (see e.g. Matarić and Michaud 2008). The consequences of learning and emergence of behaviors may be described in the product specification when a robot leaves the factory, but it is not clear how developers or users of this robot can be held responsible for its future actions. Sparrow (2007) concludes that military robots with this kind of unpredictable behavior simply should not be deployed in war scenarios since no one can be held responsible and there is no way to really punish the robots if they misbehave. While this may be a reasonable conclusion today, advanced learning capability will not only make it harder to blame developers and users of robots, but will also make it more reasonable to assign responsibility to the robots. If a robot learns and changes behavior as a result of praise and blame it receives,[6] it may actually make sense to "punish" the robot. It may even fulfill Aristotle's requirement for a morally responsible agent: worthy of praise or blame for its actions. We argue that the ability to learn (a property already covered by the concept autonomous power) will be a decisive factor for our inclination to attribute moral responsibility to robots. Another response to Sparrow's concern would be that our societies may decide to collectively share responsibility for war robots' behavior. This may be seen as an alternative if the advantages with the robots are so big that we accept the disadvantages as a calculated risk worth taking. This way of dealing with risks is already applied in a number of other areas. The way our societies support traffic is one example. Traffic is well known to involve a huge number of risks and negative consequences, including traffic accidents and air

[5] Also see https://edocs.uis.edu/kmill2/www/TheRules/ (Accessed January 3, 2012).

[6] Basic forms of this type of *reinforcement learning* is already developed and used in robotics (see e.g. Hertzberg and Chatila (2008) or Sutton and Barto (1998)), with inspiration from the "Law of Effect", a model of human and animal learning introduced by Thorndike (1911).

pollution. When people are killed by these causes, anyone is rarely held responsible for decided speed limits, or trade-offs between safety and costs when building roads. Instead, an underlying agreement is that the advantages with our traffic systems are larger than the disadvantages. The same way of reasoning could in principle be applied to unpredictable, but appreciated, autonomous war robots.

Independent of the responsibility issue, the moral quality of robots' behaviors should be seen as one of many performance measures by which we evaluate robots. We want a robot vacuum cleaner to remove dust without ruining the furniture, and we want a battlefield robot to conduct its tasks in an ethically acceptable way. Just as in the case with moral responsibility, we support a pragmatic approach focusing on subjective judgments of the robots' behavior, rather than on an objective analysis of good and bad actions. As Asaro (2006) argues, "… our overarching interest in robot ethics ought to be the practical one of preventing robots from doing harm …". Leaning against formal rules and regulations is one way to go. Ethical rules for human behavior in armed conflicts are expressed in international Laws of War, and may form a basis for ethics based control systems for robots. One corner stone of Laws of War is *Discrimination*, the principle that only combatants are legitimate targets of attack. Realization of this principle in a robot requires, among other things, advanced image analysis, analysis of other sensor data, and data fusion. Significant breakthroughs are necessary for solutions to work in real-world situations with complications such as varying light conditions and partial occlusion. A general solution would need breakthroughs also for very hard general artificial intelligence problems such as situational awareness and object recognition. Initial ideas (deliberately ignoring the mentioned technological problems) for ethics based control systems have been presented by Arkin (2009a, b). In the long run, it is likely that international Laws of War will be amended with specific requirements on the behavior of autonomous robots, thereby further motivating using such rules as basis for ethics based control systems. For a thorough overview of challenges and possibilities in construction of ethical autonomous military robots, see (Lin et al. 2008). Even the most optimistic and enthusiastic proponents of ethical battlefield robots realize the extremely hard technological requirements for universally working solutions. *Bounded morality* (Allen et al. 2006) is one approach to make ethical military robots possible already with today's technology. The idea is to limit usage of the robots to very narrow and specific situations such as taking a building, and not for the full spectrum of combat. In this way, ethical considerations are greatly simplified. In the case of Phalanx, the ability to fire autonomously is geometrically limited to a predefined "kill box" in which no civilians are allowed to be, thereby greatly simplifying (or even eliminating) the Discrimination problem. Another way to deal with the lack of universally working solutions is to keep the human "in the loop", by requiring human acknowledgement before an otherwise autonomous robot may activate its weapons. This kind of supervisory control (Sheridan 1992) is commonly used for many types of semi-autonomous robots. It should be noted that the long-term goal of the military is to have the human rather "on the loop" than "in the loop", thereby only monitoring the automatic activation of weapons (U.S. Air Force 2009, p. 41). Furthermore, as advanced mobile combat robots will be developed, bounded morality and kill boxes will be of limited value, since robots that move around freely would have to make decisions on how to act in very many different, and even previously unseen, situations. How to design ethics based control systems should therefore be discussed already now. From a consequentialist view, it would indeed be highly immoral to develop autonomous robots capable of deciding on and performing acts involving life and death, without including some kind of moral framework. This becomes even more urgent if we take into account that some of these robots may be viewed as partly morally responsible for their actions. Initially, these moral frameworks will be primitive and insufficient compared to the advanced moral laws and rules we use to judge human behavior in war. However, the alternative is to leave the stage free for purely technical considerations and limited possibilities to specify, standardize, control and criticize the behavior of battlefield robots.

Acknowledgments The author would like to thank several anonymous reviewers for their highly valuable comments and suggestions to this and earlier versions of the paper.

References

ABIresearch. (2011). *Military robot markets to exceed $8 billion in 2016*. Retrieved June 15, 2012, from http://www.abiresearch.com/press/3616-Military+Robot+Markets+to+Exceed+%248+Billion+in+2016.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems*, 12–17, July/August.

Aristotle. (1985). *The Nicomachean Ethics* (Terence Irwin, Trans.). Hackett Publishing Co, 1985.

Arkin, R. C. (2009a). *Governing lethal behavior in autonomous robots*. London: Chapman & Hall/CRC.

Arkin, R. C. (2009b). Ethical robots in warfare. IEEE Technology and Society Magazine, *28*(1), 30–33, Spring 2009.

Asaro, P. M. (2006). What should we want from a robot ethic? *IRIE International Review of Information Ethics*, *6* (12/2006).

Bechtel, W. (1985). Attributing responsibility to computer systems. *Metaphilosophy, 16*(4), 296–306.

Bone, E., & Bolkcom, C. (2003, April). *Unmanned aerial vehicles: Background and issues for congress*. Retrieved January 3, 2012, from https://www.policyarchive.org/handle/10207/1698.

Connolly, W. (1974). *The terms of political discourse*. Princeton: Princeton University Press.

Dennett, D. C. (1973). Mechanism and responsibility. In T. Honderich (Ed.), *Essays on freedom of action*. Boston: Routledge & Keegan Paul.

Dennett, D. C. (1997). When HAL kills, who's to blame? computer ethics. In D. G. Stork (Ed.), *HAL's Legacy: 2001′s computer as dream and reality*. Cambridge: MIT Press.

Dodig-Crnkovic, G., & Persson, D. (2008). Sharing moral responsibility with robots: A pragmatic approach. In A. Holst, P. Kreuger, & P. Funk (Eds.), *10h Scandinavian Conference on Artificial Intelligence SCAI 2008* (Vol. 173). Frontiers in Artificial Intelligence and Applications.

Eshleman, A. (2009). Moral responsibility, the stanford encyclopedia of philosophy (Winter 2009 Edition). In E. N. Zalta (Ed.), Retrieved January 21, 2012, from http://plato.stanford.edu/archives/win2009/entries/moral-responsibility/.

Franklin, S., & Graesser, A. (1997). *Is it an agent, or just a program?: A taxonomy for autonomous agents* (pp. 21–35). Berlin: Intelligent Agents III.

Friedman, B. (1990). *Moral responsibility and computer technology*. Erin document reproduction services.

Friedman, B., & Millett, L. (1995). It's the computer's fault—reasoning about computers as moral agents. In *Conference companion of the conference on human factors in computing systems* (pp. 226–227). Denver, CO.

Friedman, B., & Millett, L. (1997). Reasoning about computers as moral agents: A research note, in human values and the design of computer technology. In B. Friedman (Ed.), Stanford/New York: CSLI Publications/Cambridge University Press.

GlobalSecurity. (2012). *TALON small mobile robot*. Retrieved January 22, 2012, from http://www.globalsecurity.org/military/systems/ground/talon.htm.

Grossman, N. (2007). Rehabilitation or revenge: Prosecuting child soldiers for human rights violations. *Georgetown Journal of International Law*, *38*, 323–362.

Hertzberg, J., & Chatila, R. (2008), AI reasoning methods for robotics. In *Springer handbook of robotics* (pp. 207–223).

Hildebrand, A. (2009, March). *Samsung Techwin's latest: A killing robot, info4 4 SECURITY*. Retrieved January 21, 2012, from http://www.info4security.com/story.asp?storycode=4121852.

Hinds, P., Roberts, T., & Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human-Computer Interaction, 19*, 151–181.

iRobot. (2012). Retrieved January 22, 2012, from http://www.irobot.com/gi/ground.

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

Kim, T., & Hinds, P. J. (2006). Who should i blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *Proceedings of RO-MAN'06* (pp. 80–85).

Lin, P., Bekey, G., & Abney, K. (2008). *Autonomous military robotics: Risk, ethics, and design, a US department of defense office of naval research-funded report*. Retrieved June 16, 2012, from http://ethics.calpoly.edu/ONR_report.pdf.

Matarić, M. J., & Michaud, F. (2008). Behavior-based systems. In B. Siciliano, & O. Khatib (Eds.), *Springer handbook of eobotics* (pp. 891–909). Springer.

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

Miller, K. W. (2011). Moral responsibility for computing artifacts: "The rules". *IT Professional, 13*(3), 57–59.

Moon, Y., & Nass, C. (1998). Are computers scapegoats? Attributions of responsibility in human-computer interaction. *International Journal of Human-Computer Studies, 49*(1), 79–94.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, & Cybernetics, 30*(3), 286–297.

QineticQ. (2012). *MAARS—Modular Advanced Armed Robotic System*. Retrieved January 22, 2012, from http://www.qinetiq-na.com/products/unmanned-systems/maars/.

Raytheon. (2009). *News release*. Retrieved January 22, 2012, from http://www.raytheon.ca/rtnwcm/groups/rcl/documents/content/rcl_archive_phalanx_release.pdf.

Riedel, F. W., Hall, S. M., Barton, J. D., Christ, J. P., Funk, B. K., Milnes, T. D., et al. (2010). Guidance and navigation in the global engagement department. *Johns Hopkins APL Technical Digest*, *29*(2).

Samsung. (2012). *SGR-1*. Retrieved January 22, 2012, from http://www.samsungtechwin.com/product/product_01_02.asp.

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

Singer, P. W. (2009a). *Wired for war—The robotics revolution and 21st Century conflict*. Penguin.

Singer, P. W. (2009b). Military robots and the laws of war. *The New Atlantis*, Winter 2009.

Singer, P. W. (2009c). Wired for war? Robots and military doctrine. *JFQ: Joint Force Quarterly,* 2009 1st Quarter, *1*(52), 104–110.

Sofge, E. (2009). *America's Robot Army: Are Unmanned Fighters Ready for Combat?* Retrieved January 3, 2012, fromhttp://www.popularmechanics.com/technology/military/robots/4252643.

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

Stahl, B. C. (2004). *Responsible management of information systems*. Hershey: Idea-Group Publishing.

Strawson, P. F. (1974). *Freedom and resentment, in freedom and resentment and other essays*. London: Methuen.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge: MIT Press.

Thorndike, E. L. (1911). *Animal Intelligence* (2nd ed.). New York: Hafner. Transaction Publishers, 2000.

U.S. Air Force. (2006). 'Reaper' moniker given to MQ-9 unmanned aerial vehicle. In *The official Web site of U.S. Air Force*. Retrieved January 3, 2012, from http://www.af.mil/news/story.asp?storyID=123027012&page=2.

U.S. Air Force. (2009). *Unmanned Aircraft Systems Flight Plan 2009–2047*. Retrieved January 3, 2012, from http://www.globalsecurity.org/military/library/policy/usaf/usaf-uas-flight-plan_2009-2047.pdf.

U.S. Navy. (2011). *MK 15—Phalanx Close-In Weapons System (CIWS), United States Navy Fact File*. Retrieved June 14, 2012, from http://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=487&ct=2.

Walzer, M. (2006). *Just and unjust wars: A moral argument with historical illustrations*. Basic Books.

Wezeman, S. (2007). UAVS and UCAVS: Developments in the European Union. *European Parliament*, October, 2007. Retrieved January 3, 2012, from http://www.europarl.europa.eu/activities/committees/studies/download.do?file=19483.

Yamauchi, B. (2004). PackBot: A Versatile platform for military robotics. In *Proceedings of SPIE Vol. 5422: Unmanned ground vehicle technology VI, , Orlando, FL.*

Yamauchi, B., Pook, P., & Gruber, A. (2002). Bloodhound: A semi-autonomous battlefield medical robot. In *Proceedings of the 23rd Army Science Conference, U.S. Army, Orlando, FL.*

Young, I. M. (2010). Responsibility and global labor justice. In G. Ognjenovic (Ed.), *Responsibility in context: Perspectives*. Springer.