Recommendations for Future Development of Artificial Agents

Deborah G. Johnson and Merel Noorman

Introduction

A set of technologies, loosely referred to as 'artificial agents', is becoming more pervasive and more powerful in the current computing landscape. . All artificial agents are built on a computational foundation though some are purely computational, e.g., Internet bots, search engines, and others are physically embodied entities with computational decision-making components, e.g., robots, unmanned aerial vehicles (UAVs), and autonomous cars. The noteworthy feature of artificial agents – the feature that leads to the artificial agent label – is their capacity to operate autonomously. At some level or to some degree, artificial agents operate independently from the humans who design and deploy them. They are agents in the sense that we deploy them to perform tasks on our behalf and often these tasks involve learning and decision-making. Since humans previously performed many of these tasks, we mark the difference, that is, the machine performance of these tasks, by referring to them as 'artificial'.

Responsibility issues are prominent in the discourse on artificial agents. Much attention has been given to the possibility that artificial agents might develop in ways that will make it impossible to hold humans responsible for their behavior. We believe that this concern misconstrues the situation and distracts attention from more important and more urgent issues. Rather than lamenting the possibility of artificial agents for which no one can be responsible, attention should be focused on how to develop artificial agents so as to ensure that humans can be responsible for their behavior. This involves attending to: the optimal distribution of tasks among human and non-human (machine) components of artificial agent systems; appropriate designations of responsibilities to the humans operating in the system; and development and implementation of responsibility practices to support the assignments of tasks and responsibilities.

As part of a National Science Foundation funded project (Ethics for Developing Technologies: An Analysis of Artificial Agent Technology, NSF Award # 105845), we developed a set of recommendations for the future development of artificial agents.[1] An early version of the recommendations was presented to a small group of experts in the fields of robotics, computer science, philosophy, ethics, law, and policy.[2] Based on the

[1] Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[2] The draft recommendations were presented as part of a workshop, Anticipatory Ethics, Responsibility and Artificial Agents, held at the University of Virginia on January 24-25, 2013.

feedback received from these experts, the recommendations were revised. We present the final recommendations here along with our rationale for putting them forward.

Background: Addressing concerns about responsibility

As already mentioned, much attention has been given to the question whether it will be possible to hold humans responsible for the behavior of artificial agents of the future. Matthias (2004) refers to this problem as 'the responsibility gap'. Artificial agents can be programmed to learn as they operate, and because of this learning, the fear is that no humans – even those who program the agents – will be able to understand how some artificial agents arrive at their decisions. Hence, no human can fairly be held responsible for what the artificial agents do.

Although the problem seems plausible as described, the responsibility gap is generally framed as an inevitable outcome of the complexity of the technology; this can be seen, for example, in Matthias' presentation of the problem. However, if a responsibility gap occurs in the future, it will not be because of the complexity of the technology; rather, it will be because of human decisions to deploy technologies without knowing how they will behave. The issue is not whether humans will understand how the technology operates. At some level, humans will understand; they just won't be able to directly control or fully predict how the agents will behave in specific circumstances. The major concerns in this situation will be reliability and safety. Will the agents have a level of reliability and safety appropriate for the tasks performed? Humans will make these decisions. So, if agents for which no human can fairly be said to be responsible are released, that will be because humans will have decided both to trust particular learning algorithms and to do without social practices that assign responsibility to human decision makers.

Scholars and researchers have taken a number of approaches to address the responsibility issues posed by artificial agents. Several have offered rules or informal laws for those who design and use artificial agents to encourage the clear allocation of responsibility. For example, Miller initiated and led a collective effort to develop a set of rules for 'moral responsibility for computer artifacts' (Miller, 2011; Grodzinsky, et. al., 2012). The rules state that the people that design, develop and deploy computational artifacts have a shared responsibility. The rules also indicate that people who knowingly design, develop and deploy these technologies can only do so responsibly when they take into account the sociotechnical systems in which the artifact is embedded. Moreover, they specify that people should not deceive users about the artifact or its foreseeable effects. Similarly, Murphy and Woods (2009) developed three laws of responsible robotics, intended as alternatives to Asimov's three rules from *I, Robot*. Central to Murphy and Woods' laws is the idea that robots should be designed to be responsive to humans. For example, Murphy and Woods offer as a first law that "A human may not deploy a robot without the human–robot work system meeting the highest legal and professional standards of safety and ethics".

Other scholars have focused on the design of technology as an important place to address responsibility issues (Lokhorst and van den Hoven, 2012). In this vein, some try to solve the problem by programming agents to be ethical (Anderson and Anderson, 2010) or programming them to adhere to rules constituting ethical practices such as the rules of engagement in war (Arkin, 2009). As early as 1994, Weld and Etzioni (1994) proposed the idea of programming agents to avoid doing harm. Other design approaches take the approach of designing systems to facilitate humans to act responsibly as they control and interact with a system (Cummings and Dipl, 2009; Cummings et al, 2010).

Non-technical approaches have focused on adjusting and developing regulations and legal frameworks for the development and use of these technologies (Asaro, 2012). In the case of military robots, for instance, several scholars and non-governmental organizations have argued for strict regulation of these technologies and even a prohibition on the development of artificial agents that take the human out of the decision-making loop (Sharkey, 2008; Sparrow 2009; Human Rights Watch, 2012).

In what follows, we offer a set of recommendations that differ from those mentioned earlier by targeting the many different actors involved in the development of artificial agents. We have in mind not just the developers and users, but policy makers, the public, managers, financiers, journalists, manufacturing companies and others who influence the development of artificial agents and the ways in which they are understood. Addressing the responsibility issues posed by artificial agents involves attention to social conceptions and social practices as well as the design of artifacts. Even if developers design a technology to encourage and facilitate responsible behavior of human users, social practices still must be developed to reinforce or facilitate the individuals and groups acting in or with the system to recognize and fulfill their responsibilities.

The Recommendations and their Rationale

One overarching principle precedes and permeates all of our recommendations: responsibility for the behavior of artificial agents resides with human beings. We believe this principle should be explicitly and publicly recognized and should inform and guide the development of artificial agents. We emphasize this principle in response to rhetoric suggesting that artificial agents for which no human can be responsible will inevitably be developed or suggesting that the concept of responsibility can be extended to artificial agents themselves. Of course, we are not the first to emphasize human responsibility. Others including artificial agent developers have recognized this principle. For example, in their paper on responsible agent behavior, Mamdani and Pitt (2000) explain that "…the interaction between agent and owner must be tightly coupled in the sense of an identifiable human taking responsibility for an agent's behavior." "Tight coupling", they write, "does not imply that an agent's every action needs the owner's sanction, but that the owner assumes responsibility for all possible behavior of an agent while it acts autonomously as its owner's surrogate" (p. 28).

Keeping this principle front and center goes hand in hand with recognizing the contingency of technological development and resisting the presumption that technological development is merely an inevitable outcome of what nature allows. What developing technologies will come to look like in the future, how they will operate, how they will be used, and who will be responsible for them are all matters contingent upon negotiations among many different actors (Bijker, Hughes, and Pinch, 1987). Engineers, researchers, and users have a strong influence on technological developments, as do financiers, lawyers, policy makers, the media, and the public. What gets designed, adopted, and used is, thus, the result of many human actors making decisions that accumulate to produce an artifact as well as a set of social practices that together constitute a sociotechnical system.

Recommendation 1:
Artificial agents should be understood to be sociotechnical systems or networks consisting of artifacts and social practices organized to accomplish specified tasks through their interactions.

We can understand artificial agents to be merely artifacts (software), but to do so is to abstract the artifactual component from the system in which it is embedded. Without the system the software has no functionality, no meaning, no effect. Consider, for example, that a bot is nothing without the Internet, the operators that maintain the Internet, users who receive the results of bot operations, and so on. Or consider an autonomous car. The physical car functions as part of a system involving other artifacts (e.g., roads, stop lights, gas pumps) and humans. Autonomous cars come into being because of human decision-making and behavior, and they operate only in combination with human behavior. Humans design and build the cars and put them on the road. During the development phase, humans monitor the behavior of the cars, making adjustments as needed. Humans decide when the autonomous cars are ready for road testing and humans will decide when they are safe enough for commercial release. Moreover, how autonomous cars operate once they are introduced on public roads is as much dependent on their many different technical components, as it is on a variety of social practices that ensure that the cars are operational as well as safe. Yes, intelligent algorithms, sensors and vast amounts of data enable the car to drive without a human driver manually controlling it, but the car's ability to navigate the roads also depends on a wide range of human behavior including legal and social norms that have been put in place through social processes and inculcated in human drivers. In order to understand why an autonomous car does what it does or when it fails to function as expected, we have to look at the car as a sociotechnical system.

Failure to recognize artificial agents as sociotechnical systems renders the role of human actors in technological outcomes invisible; it invites the idea that no humans are responsible for those outcomes. Taking a sociotechnical perspective on artificial agents brings into view the social as well as technical components that constitute the technology, so that attention can be given to how the two work together.[3] This perspective facilitates

[3] See for example (Verbeek 2000, Chap. 8; Verbeek 2011).

developers and engineers as well as policy-makers and managers in seeing the multiple actors involved in artificial agent development and operation. It facilitates recognition of the interactions between humans and agents and the different levels of responsibility. It helps to anticipate uncertainties in the development and implementation process.

To illustrate how a sociotechnical perspective can help in reasoning through how a new technology can affect responsibility in a network of human and non-human actors, think of a healthcare robot that can autonomously find its way around a hospital to deliver telemedicine care (Tsui and Yanco, 2007; Ackerman 2013). Such robots operate alongside and in interaction with hospital employees, patients and visitors. Of course, developers and engineers have a responsibility to ensure that the technology operates according to its technical specifications and that it does not pose any danger to the people interacting with it. At the same time, users and operators have a responsibility to use the technology as it is intended to be used. A sociotechnical perspective, however, also draws attention to the responsibilities of the other human actors that are involved in the development and operation of these robotic systems. For instance, engineers on the design team were responsible for making sure that the technology was adequately tested; the hospital is responsible for training those who operate and maintain the robot; during the purchase process, hospital representatives had to make sure the robot would not operate in a way that conflicts with other organizational procedures and practices within the hospital; and so on.

Recommendation 1 is not just in harmony with our overarching principle; it supports recognition of the principle. When those involved fail to see that artificial agents are sociotechnical systems, they are more likely to overlook the responsibilities of human actors other than operators and users - those human actors that create the conditions for the operation of the technology. This changes when the technical components that constitute artificial agents are understood in terms of their role in a broader network of human and non-human components. Failure to recognize this may lead to a failure to see that part of the development of a technology is designing the way humans will operate in or interact with the system and interact with artifactual components. In the case of artificial agents, this means, among others, those who decide when and where to deploy the agents, those who control the artifact in operation, and those who monitor and maintain the agents.

Recommendation 2:
Responsibility issues should be addressed when artificial agent technologies are in the early stages of development.

Responsibility issues often seem to be addressed (if at all) only when a technology is introduced in real-world settings. Google's on the street experiments with their autonomous car quickly generated a debate about the applicability of existing legal frameworks and liability laws in California and Nevada (Pinto 2012). The technology was already developed to a significant extent and decisions about how the car should behave had been made along the way. Before these experiments, discussions about responsibility

involving autonomous cars were relatively few and did little to influence the design of the car.

This is not uncommon for new technologies. In the early stages of development the focus tends to be on technical feasibility. UAVs, for instance, have been in development for over fifty years. Yet, the discussions about responsibility have only recently gained momentum especially because of the use of drones in military operations. To be sure, issues of responsibility with regard to autonomous robots have been raised early on. The handful of roboticists who work on designing robots to be ethical (Arkin, 2009; Anderson and Anderson, 2010) are an illustration of this. Yet, it was only after these technologies were introduced on a larger scale that the debate really took off.

The problem of addressing ethical issues after a technology is introduced and put into use is that the design decisions that have already been made may preclude or constrain the possibilities for particular kinds of responsibility assignments. It is also possible that responsibilities for various aspects of a new system may be assumed or given cursory attention during one phase or another, e.g., testing, rather than being assigned explicitly and carefully with an eye to the users or the context of use. This makes the issues harder to address and may do damage to acceptance of the new technology.

An important trend in recent scholarship on ethics and technology is to identify and analyze the ethical issues raised by new technologies while these technologies are still in the early stages of development (van de Poel, 2008; Verbeek, 2011; Brey, 2012a). Some refer to this endeavor as anticipatory ethics (Carter, et. al., 2009; Johnson, 2011) or anticipatory technology ethics (Brey, 2012a, 2012b). Value sensitive design (VSD) and responsible innovation (van den Hoven & Manders-Huits, 2009; von Schomberg, 2011) are examples of such approaches. Such anticipatory ethics approaches generally build on a sociotechnical perspective. They analyze the sociotechnical processes that underlie the development of a new technology in order to intentionally address ethical issues before the technology has stabilized.

An example of how concerns about responsibility can be part of early development stages is when designers build in monitoring systems to record information on the operation of a vehicle with an eye to determining the chain of events in cases of accidents. Here the monitoring system may be used not just to trace back what happened but also to decide who is accountable for an accident or failure. Another example is user interface design that can be designed to facilitate or constrain responsible operation of a UAV or other agent (see, for example, Cummings, 2004).

In emphasizing the value of addressing responsibility issues early on, Recommendation 2 helps to ensure that responsibility issues do not fall through the cracks and decreases the possibility that artificial agents that defy human control will be implemented. Intentionally bringing ethical concepts and concerns into the early stages of technological development has the potential more broadly to lead to better technologies. We are more likely to get

technologies that are aligned with moral notions and less likely to get technologies that violate moral sensibilities and principles. We are more likely to eschew lines of research or designs that ultimately would have been rejected by consumers, users or the public.

Recommendation 3:
Claims about the autonomy of artificial agents should be judiciously and explicitly specified.

In the discourse on artificial agents, a great deal of attention is given to the autonomy of artificial agents. A number of scholars have devoted attention to futuristic visions, similar to those found in popular media, in which artificial agents will have or even exceed the capacities of human beings (see, for example, Sparrow, 2004). Such visions suggest that future agents will be increasingly autonomous, operating entirely independent of direct human control. This prospect underlies many of the concerns about whether or not humans can still be held responsible. Moreover, the idea of increasingly autonomous artificial agents leads to speculations and discussions about the moral rights and moral standing of these agents (Whitby, 2008; Hellström, 2012).

This speculation and discussion illustrates how imaginative future visions of technological development can be misleading. Computer scientists use the concept of autonomy to explain how new technologies would work and to provide a goal to strive for. Some computer scientists use autonomy to refer to high-level automation. Some use it to refer to the capacity that some artificial agents have to navigate in environments by relying on machine models of the environment. Yet others see machine autonomy simply as responsiveness to an environment wherein sensors provide input and artificial agents are programmed to respond to those inputs. [See Johnson and Noorman, 2014 for a discussion of different conceptions of autonomy.] However valuable, these interpretations of autonomy are distinct from conceptions of autonomy in moral philosophy or daily life, where autonomy is intertwined with notions of free will, responsibility and intentionality. Failing to recognize the differences between these conceptions of autonomy can lead to confusion about how artificial agents actually work and to unjustified inferences about responsibility, for example, to the unjustified conclusion that entities with machine autonomy can, themselves, be responsible or that humans are less responsible for the behavior of such entities.

As the various conceptions of autonomy suggest, claims about the autonomy of artificial agents have to be understood within the context in which they originate. For example, in developing policies and regulation for the use of autonomous cars on public roads, it is not enough to simply assume that the autonomy of these cars is self-explanatory. Autonomous cars may be driverless cars or cars with some autonomous components, such as a self-parking function. The difference is important, among other things, for policy making since policies would be different for each (Pinto 2012). Similarly, the prospect of increasingly autonomous military robots draws attention away from the activities of operators, technicians, and other human actors involved in the operation of these robots. The latest generation of UAVs are said to be autonomous because they can take off from and land on

an aircraft carrier on their own. However, this kind of autonomy is very far off from the kind of autonomy that the media and popular science articles predict.[4] That is, taking-off, landing and navigating is quite different from independently making targeting decisions in active combat.

Another problem with underspecified and context-less conceptions of autonomy is that they obfuscate the relationship of control between human actors and the technology. When the autonomy of a UAV refers to its ability to take off, land, and navigate between waypoints without human intervention under 'normal conditions', the operator may still be required to monitor the performance of the aircraft when the conditions are irregular. Such an autonomous technology has a human operator, an operator that must be in a position to exercise responsibility, e.g. she has to be continuously aware of and able to understand the status of the aircraft. Specifying and contextualizing the meaning of 'autonomous' leads, thus, to paying attention to assumptions about the distribution of tasks and control among human and non-human components.

Recommendation 3 is not just important for developers who typically present new technologies to funders, users, and the public and may have an incentive to use hyperbole; it is important for all who communicate to the public. This includes policy makers and legal experts, as well as researchers, scholars, and journalists. Keeping claims about the autonomy of artificial agents in context can contribute to a more informed public and the possibility of better public input to the future development of artificial agents.

Recommendation 4:
Responsibility issues involving artificial agents can best be addressed by thinking in terms of responsibility *practices*.

Part of the problem in addressing responsibility issues raised by complex technologies is that the conventional moral notions of responsibility are difficult to apply to humans acting with technologies. These notions place an emphasis on the relationship between an individual's actions and capacities, and the outcome of those actions and capacities ignoring the contribution made by a technological environment. In technological environments, the individual may not have full control of the outcomes and may have had only a partial understanding of where their individual action would lead. The technological contribution exacerbates the challenges of locating responsibility. When something untoward happens, such as a drone killing an innocent person, tracing back what went wrong and identifying what or who is responsible can be a daunting task, though that does not mean that no one is responsible.

---

[4] See, for example, http://www.aljazeera.com/indepth/features/2013/04/201344132214594527.html (last modified: 08 Apr 2013 12:32).

Because of the complexities of responsibility ascription, it is useful to think about responsibility as a set of practices (Noorman, 2013). That is, what or who is responsible for an outcome depends on the responsibility practices in the context at issue. These practices are the established ways in which groups and individuals in a community understand, evaluate, and distribute responsibility. Individuals and groups come to have a responsibility or to be responsible for something as a result of social norms and shared ideas about what sort of behavior is expected in particular contexts and what consequences will follow from living up to or failing to live up to the expectations. These norms are established and reinforced in a wide variety of ways. For example, managers and colleagues inform new engineers about what their job entails and what is expected of them. Their tasks and duties may be explicitly stated in training manuals, procedures, protocols, codes of ethics and regulations, but engineers may also learn about these as they become acquainted over time with the informal norms and rules within organizations. The engineer's acceptance of a job is taken as a sign that he or she accepts the assignment of tasks and duties that go with the job. When an engineer fails to live up to specified responsibilities he or she can be reprimanded by colleagues or superiors through informal conversations or through more formal processes, e.g., annual evaluations. In high profile cases such as occurred with the Challenger disaster and the Y2K problem, the public may demand an explanation, an accounting from the profession. These activities all contribute to shared norms of responsibility; they constitute responsibility practices.

Responsibility practices involve formal and legal frameworks as well as informal social conventions and are usually a mixture of both. A soldier is responsible for following orders given by a superior officer; this responsibility is reinforced by military culture and daily practice but it is also formally expressed in military codes, codes that have been refined in military court decisions. In other contexts responsibility practices may be contested, for example, informally if a search engine bot delivers pornographic material in response to a search word entry; some would blame the searcher, others might blame the search engine designers or the website owners.

The notion of responsibility practices places the focus on the shared understanding of what is expected, what is owed, and what are likely consequences of failure within a sociotechnical network. Using this notion involves examining what constitutes responsibility in a particular network and why practices have been constituted as such: what are blame- or praiseworthy actions; when is one required to account for one's behavior; what would cause feelings of guilt and shame, regret, and a sense of moral reckoning; and so on. Using the notion of responsibility practices draws attention to the formal and informal mechanisms and strategies, such as legal procedures, organizational rules, and social norms that promulgate and enforce responsibility.

Once established, responsibility practices are continuously challenged and renegotiated. New technologies, for instance, tend to change the way people do things, how they relate to each other and, thus, how responsibilities are distributed and acted on. For example, the introduction of UAVs has led to a range of new procedures and protocols in military

operations (Noorman 2013). Pilots had to learn new skills, commanders had to develop new strategies, and decision-making processes have undergone significant changes. This has meant in turn that new practices have had to be developed and negotiated for assigning and ascribing responsibility, not only within military organizations but outside as well. Increasingly prominent discussions in the mainstream media about the ethics of drone warfare show that the U.S. is still in the process of developing new norms, policies and laws that govern the use of these unmanned systems.

In order to address responsibility concerns involving artificial agents, a careful analysis of responsibility practices and how they shape and are shaped by the design of a particular technology is required. When responsibility practices are considered in the early stages of development (as indicated with Recommendation 1), then policy and practices that assign and specify who is responsible for what can be put in place and enhanced as the technology reaches completion. It is important to take into consideration how the relationships between the relevant actors take shape, how ideas about obligations, duties and accountability evolve, and what mechanisms and strategies are put in place to enforce these responsibilities on multiple levels. It also means that we have to create the appropriate conditions for people to be responsible: institutional mechanisms, law and policy, etc.

Conclusion

The overarching principle and four recommendations work together to constitute a framework for thinking about artificial agent technology (and other emerging technologies) that enables consideration of issues of responsibility. To effectively address the responsibility issues in artificial agent technology, it is important to address the issues early on and the issues can only be fully seen (early on or later on) when the technology is viewed as a sociotechnical system, when claims about autonomy are used judiciously and kept in context, and when responsibility is understood to be a matter of social practices rather than something in an individual.

The recommendations are broad both in the sense that they apply to many different actors and also in the sense that they imply different sorts of behavior for different actors. For example, they encourage researchers, engineers and developers to look beyond technical specifications and imagine what kind of interactions between a technology and its users can occur; they remind journalists and policy makers to be critical of abstract future visions and look more closely at the possibilities and limitations of the technology in particular contexts. The various actors thus need to reflect on the technical, social and ethical aspects of artificial agent technologies.

In facilitating consideration of issues of responsibility, these recommendations are targeted to improve technological development. Artificial agents of the future promise to perform some tasks better than humans and to make it possible for humans to do things they could not have done before. However, this will not happen unless humans continue to take

responsibility for their design and their behavior. These recommendations are put forward to that end.

References

[1] Ackerman, E. (2013). RP-VITA approved for hospital use, SUGV approved for disruptor use. IEEE Spectrum (January 21, 2013). Available at: <http://spectrum.ieee.org/automaton/robotics/medical-robots/rpvita-approved-for-hospital-use-sugv-approved-for-disruptor-cannon-use> . Last accessed January 9, 2014.

[2] Anderson, M., & Anderson, S. (2010). Robot be good. Scientific American, October 2010, 72-77.

[3] Arkin, R. C. (2009). Ethical robots in warfare. Technology and Society Magazine, IEEE, 28(1), 30-33.

[4] Asaro, P. M. (2011). 11 A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics. In P. Lin, K. Abney, and G.A. Bekey, *Robot Ethics: The Ethical and Social Implications of Robotics*, pp. 169-186.

[5] 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.

[6] Brey, P.A.E. (2012a). Anticipatory Ethics for Emerging Technologies. NanoEthics, 6(1), 1-13.

[7] Brey, P.A.E. (2012b). Anticipating ethical issues in emerging IT. Ethics and Information Technology, forthcoming.

[8] Carter, A., Bartlett, P., & Hall, W. (2009). Scare-Mongering and the Anticipatory Ethics of Experimental Technologies. American Journal of Bioethics, 9(5), 47-48.

[9] Cummings, M.L., Clare, A. and Hart, C. (2010). The Role of Human-Automation Consensus in Multiple Unmanned Vehicle Scheduling.
Human Factors: The Journal of the Human Factors and Ergonomics Society. 52(1), pp 17-27.

[10] Cummings, M. L. & Dipl, S.B. (2009). Collaborative Human–Automation Decision Making. Springer Handbook of Automation, pp 437-447.

[11] Cummings, M. L. (2004). Creating moral buffers in weapon control interface design. *Technology and Society Magazine, IEEE*, *23*(3), 28-33.

[12] Grodzinsky, F.S., Miller, K., and Wolf, M.J. (2012). Moral responsibility for computing artifacts: "the rules" and issues of trust. SIGCAS Comput. Soc., Vol. 42, No. 2 (December 2012), 15-25.

[13] Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and information technology*, *15*(2), 99-107.

[14] Human Rights Watch and International Human Rights Clinic (2012). Losing humanity: The case against killer robots. Report. Available at: <http://www.hrw.org/reports/2012/11/19/losing-humanity-0>. Last accessed March 20, 2013.

[15] Johnson, D. (2011). Software agents, anticipatory ethics, and accountability. In Marchant, G.E. B.R., Allenby, and J.R. Herkert (Eds.). The growing gap between emerging technologies and legal-ethical oversight (pp. 61-76). [The international library of ethics, law and technology 7.] Dordrecht: Springer.

[16] Lokhorst, G. J., & van den Hoven, J. (2012). 9 Responsibility for Military Robots. Robot Ethics, 145.

[17] Mamdani, E., & Pitt, J. (2000). Responsible agent behavior: a distributed computing perspective. *Internet Computing, IEEE*, *4*(5), 27-31.

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

[19] Miller, K. W. (2011). Moral Responsibility for Computing Artifacts. IT Professional, 13(3), 57-59.

[20] Murphy, R.R. and Woods, D.D. (2009). "Beyond Asimov: The Three Laws of Responsible Robotics," Intelligent Systems, IEEE , 24(4), (July-August, 2009):14-20.

[21] Noorman, M., & Johnson, D. G. (2014). Negotiating autonomy and responsibility in military robots. *Ethics and Information Technology*, *16*(1), 51-62.

[22] Noorman, M. (2013) Responsibility Practices and Unmanned Military Technologies. Science and Engineering Ethics, Online First, DOI: 10.1007/s11948-013-9484-x

[23] Pinto, C. (2012). How Autonomous Vehicle Policy in California and Nevada Addresses Technological and Non-Technological Liabilities. Intersect: The Stanford Journal of Science, Technology and Society, [S.l.], v. 5, jun. 2012. Available at: <http://ojs.stanford.edu/ojs/index.php/intersect/article/view/361>. Last accessed April 22 2013.

[24] Sharkey, N. (2008). Cassandra or false prophet of doom: AI robots and war. IEEE Intelligent Systems, 23(4), 14-17.

[25] Sparrow, R. (2009). Building a Better WarBot: Ethical issues in the design of unmanned systems for military applications. Science and Engineering Ethics, 15(2), 169-187.

[26] Sparrow, R. (2004). The turing triage test. *Ethics and Information Technology*, *6*(4), 203-213.

[27] Tsui, K. M. and Yanco, H. A. (2007). Assistive, Rehabilitation, and Surgical Robots from the Perspective of Medical and Healthcare Professionals. In Proceedings of AAAI Workshop on Human Implications of Human-Robot Interaction. Available at: http://www.cs.uml.edu/~ktsui/papers/2007-AAAI-Tsui.pdf. Last accessed January 9, 2014.

[28] Van de Poel, I. (2008). How should we do nanoethics? A network approach for discerning ethical issues in nanotechnology. Nanoethics, 2, 25-38.

[29] Van den Hoven, J. & Manders-Huits, N. (2009). Value-sensitive design. In Olsen, J.K.B, Pedersen, S.A., & Hendricks, V.F. (Eds.), A Companion to the Philosophy of Technology. Wiley-Blackwell.

[30] Von Schomberg, R. (Ed.) (2011). Towards Responsible Research and Innovation in the Informationand Communication Technologies and Security Technologies Fields. Luxembourg, Publications Office of the European Union.

[31] Verbeek, P. (2000). De Daadkracht der Dingen. Amsterdam: Boom.

[32] Verbeek, P. (2011). Moralizing Technology. Chicago: The University of Chicago Press.

[33] Weld, D., & Etzioni, O. (1994, October). The first law of robotics (a call to arms). In *AAAI* (Vol. 94, pp. 1042-1047).

[34] Whitby, B. (2008). Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents. Interacting with Computers, 20(3), 326-333.