Principles for the Future Development of Artificial Agents

Deborah G. Johnson
Science, Technology and Society Program
University of Virginia
Charlottesville, VA, USA
dgj7p@virginia.edu

Merel Noorman
eHumanities group
Royal Netherlands Academy for Arts and Sciences
Amsterdam, The Netherlands
merel.noorman@ehumanities.knaw.nl

***Abstract*—A survey of popular, technical and scholarly literature suggests that autonomous artificial agents will populate the future. Although some visions may seem fanciful, autonomous artificial agents are being designed, built, and deployed in a wide range of sectors. The specter of future artificial agents – with more learning capacity and more autonomy – raises important questions about responsibility. Can anyone (any humans) be responsible for the behavior of entities that learn as they go and operate autonomously? This paper takes as its starting place that humans are and always should be held responsible for the behavior of machines, even machines that learn and operate autonomously. In order to prevent evolution to a future in which no humans are thought to be responsible for the behavior of artificial agents, four principles are proposed, principles that should be kept in mind as artificial agents are developed.**

Keywords—artificial agent; responsibility; autonomy;

I. INTRODUCTION

A survey of popular, technical and scholarly literature suggests that part of the future is already set. Autonomous artificial agents will populate the future. Some will be invisible, computational devices (bots) embedded in human constructions and natural environments. Others will be visible and embodied, e.g., robots, unmanned vehicles. They will perform rudimentary as well as complicated and critical decision making tasks. Robots will serve as warfighters, healthcare providers, factory workers, domestic servants, and personal companions [4, 5]. Embodied or not, they will make speed of light transactions, run nuclear power plants, control transportation, and may even make end-of-life decisions. As some would have it, some of the robots will have moral standing [7]; they will have rights to be treated and not treated in certain ways.

Although some visions may seem fanciful, streams of research are underway that are targeted to make these visions reality. Artificial agents that are autonomous in certain respects are being designed, built, and deployed in a wide range of sectors. Computer scientists, engineers, and philosophers have even begun to develop software to give robots the capacity to make moral decisions and to behave ethically [1,2,3]. Others, drawing on research about how people respond to computers and robots, are developing robots that look and act like humans (humanoid robots) [8].

The specter of future artificial agents raises important questions about responsibility. Can anyone (any humans) be responsible for the behavior of entities that learn as they go and operate autonomously? Some have argued that as artificial agents become increasingly more autonomous, there may come a time in the future when no humans can be responsible for their behavior [6].

This paper takes as its starting place that humans are and always should be held responsible for the behavior of machines, even machines that learn and operate autonomously. Whatever technologies are developed in the future, they will be the result of negotiations among many different actors – engineers and scientists, investors, regulators, journalists, politicians, the public. Humans will make decisions about the design and the deployment of artificial agents. They will understand generally how the agents work; they just won't be able to directly control or fully predict how the agents will behave in particular circumstances. However, humans will be the ones setting the conditions and defining the criteria under which they will allow these agents to operate. The big and important issues will have to do with reliability, and safety. These issues are often at the forefront of new technologies, and artificial agents will be no exception.

Of course, the future is uncertain and it is possible that at some time in the future, people will come to think that no one can be responsible for the behavior of certain artificial agents. People may even come to believe that it makes sense to say that the machines are responsible for their own behavior. However, if this happens, it will not because the technology developed in a way that made it impossible for humans to be held responsible. Rather, it will be because humans decided to

978-1-4799-4992-2/14/$31.00 ©2014 IEEE

trust learning algorithms and to abandon practices that hold humans responsible for the behavior of the devices they design and deploy.

In order to impede evolution to a future in which no humans are thought to be responsible for the behavior of artificial agents, it is important to keep four principles in mind as artificial agents are developed. These principles should be in the forefront of the thinking of engineers and computer scientists as well as other actors who will influence the development of artificial agents.

II. THE FOUR PRINICIPLES

A. *Artificial agents should be understood to be sociotechnical systems consisting of artifacts (material objects) and social practices organized to accomplish specified tasks through their interactions.*

Taking a sociotechnical perspective on artificial agents keeps attention on agents as combinations of people and things. This perspective prevents the common error of thinking that technology is simply material objects. Artificial agents have an artifactual (material) component but their operation requires human activity. A bot performing transactions on the Internet has been designed and deployed by people. Whatever tasks an artificial agent performs, they have been created by humans to support human desires and activities. For example, a military robot operating on the battlefield has been built and programmed by human beings and operates in conjunction with people (and other machines). Whatever the connections between things and people in a sociotechnical system, ultimately one finds humans who want to achieve goals, e.g., win a battle, defeat an enemy, and gain some broader outcome.

B. *Responsibility issues should be addressed while artificial agent technologies are still in the early stages of development.*

In the early stages of technological development, the focus tends to be on technical feasibility. Ethical issues of any kind may be difficult to address because there is uncertainty as to what the technology will look like and how it will be used when it is ultimately adopted. Although this is problematic for addressing ethical issues early on, design decisions often preclude or constrain the possibilities for ethical arrangements such as responsibility assignments.

Responsibility depends on the distribution of tasks. In designing a new technology, tasks are distributed among human and non-human components. For example, in conventional automobiles, the machine components of the car perform various functions and humans provide input to these functions by pressing buttons and pedals, and steering. In autonomous cars, tasks are redistributed; some of the tasks conventionally performed by humans are moved from human to machine. Of course, some tasks continue to be performed by humans, e.g., turning the car on and off, maintaining the software, deciding when and where to go.

Even when tasks are assigned to machine components, humans have responsibilities. In the case of autonomous cars, for example, the car manufacturers, those who own and deploy the cars, and those who license the vehicles, all have responsibilities. The design of the car may make the responsibilities of various actors harder or easier to perform. For example, in the case of autonomous automobiles, built-in recording devices that monitor machine and human behavior help to sort out responsibility when accidents occur.

C. *Claims about the capabilities of artificial agents should always be framed in a particular context.*

What artificial agents are – their meaning, significance, and role in everyday life – is a matter of discursive framing. Without context, referring to artificial agents as autonomous can be misleading. 'Autonomy' is used in many different ways. Computer scientists use the concept of autonomy to describe the way certain technologies work, i.e., that, once deployed, they operate independent of human intervention. Some computer scientists use autonomy to refer to high-level automation, i.e., machines that operate independent of direct human intervention for long periods of time. They may, for example, use it to refer to the capacity that some agents have to navigate in environments by relying on machine models of the environment. Such uses of autonomy are helpful to those who are thinking, speaking and writing about artificial agents. However, such interpretations 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 the unjustified inference that entities with machine autonomy can, themselves, be responsible or that humans are less responsible for the behavior of such entities.

D. *Responsibility issues can best be addressed by thinking in terms of responsibility practices.*

Responsibility is often thought of as a simple matter; an individual is either responsible for what happened or not. However, responsibility is almost always embedded in a context in which individuals are interacting with others and with machines and devices. A variety of practices convey and reinforce norms and expectations about who is responsible for what. Individuals and groups come 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.

Responsibility practices are practices that specify, support, or reinforce assignments of responsibility and their fulfillment. For example, soldiers who press buttons deploying drones are given directions about when they are to press a button and when not, i.e., in what circumstances. Their responsibility in the button-pressing situation is reinforced by broader training in the military and by military culture. Moreover, the physical environment in which they work and the devices that they operate may constrain or facilitate the soldier's ability to fulfill specified responsibilities. A soldier's understanding of his or her responsibility is, thus, constituted through a wide range of practices – oral and written orders, broad cultural standards, knowledge of the consequences for failure, witnessing others being held responsible, etc.

Responsibility for future artificial agents will emerge from a set of practices that come to constitute the operation of such agents and it is important to recognize this so that responsibility practices can be conscientiously created.

III. CONCLUSION

These principles are targeted to improve technological development. Artificial agents of the future promise to perform certain tasks better than humans can and to make it possible for humans to do things they could not have done before. To ensure that this happens, those involved in the development of artificial agents should keep in mind that artificial agents are sociotechnical systems. They should consider issues of responsibility in the early stages of development. They should be careful in making claims about the nature and significance of artificial agents, always keeping such claims in context. Issues of responsibility should be thought of in terms of a set of responsibility practices that must be created and will become part of artificial agents (understood as sociotechnical systems).

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1058457. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Anderson, M. & Anderson, S.L. (eds.). (2011). *Machine ethics*. New York, NY: Cambridge University Press.

[2] Arkin, R.C. (2010). The Case for Ethical Autonomy in Unmanned Systems. *Journal of Military Ethics,* 9(4), 332-341.

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

[4] Chen, C. H., Weng, Y. H., & Sun, C. T. (2009). Toward the human-robot co-existence society: on safety intelligence for next generation robots. *Social Robotics*.

[5] Lin, Patrick, Keith Abney, and George A. Bekey. (2012) *Robot Ethics: The ethical and social implications of robotics*. The MIT Press.

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

[7] 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.

[8] Zhao, Shanyang. (2006) "Humanoid social robots as a medium of communication." *New Media & Society* 8(3), 401-419.