J Bus Ethics
DOI 10.1007/s10551-014-2180-1

Technology with No Human Responsibility?

Deborah G. Johnson

Introduction

A major thrust of Richard De George's book, *The Ethics of Information Technology and Business* (2003), was to draw attention to the ethical challenges for business as business practices were being reconfigured as a result of the introduction of computing and information technology. The topics on which he focused and his analysis are still relevant a decade later. Today privacy issues are pervasive. Intellectual and other kinds of property rights in electronic data and devices continue to challenge courts of law and legislative bodies. E-business is now the norm as most businesses are online in some form or another. The nature of work continues to change as new technologies are introduced; the new technologies change what workers do, when, where, and how they do it, and the extent to which they are monitored. As De George himself wrote, he was addressing a rapidly moving target, and the target—changes in the way business is done due to changes in computing and information technology—continues to move.

The starting place for De George's analysis in *The Ethics of Information Technology and Business* is a critique of what he refers to as the myth of amoral computing and information technology (MACIT). This myth, he claims, blinds us to the powerful changes taking place as a result of computing and information technology. Despite increased awareness today of many of the issues identified by De George, the business community and the public still seem to hold some version of the MACIT. That is, the belief that technological choices are amoral is fairly common even in the face of blatant evidence to the contrary, evidence indicating that technological choices have moral consequences.

De George agrees with at least part of the MACIT; he acknowledges that it "like all myths, partially reveals and partially hides reality." He writes:

> The Myth of Amoral Computing and Information Technology takes many forms. It does not hold that computing is immoral. Rather in holding that it is amoral MACIT says that it is improper, a conceptual mistake, to apply moral language and terms to computers and what they do. This much is correct. But what is false is that it is improper or a conceptual mistake to apply moral language and terms to what human beings do with computers, how they design, develop and apply them, how they manipulate and use information. (p. 6)

De George goes on to lament the lack of debate about whether we want the kind of society that accompanies computing and information technology, and he broadens the myth to include the unquestioned acceptance of technology and its seeming unstoppable progression:

> There is no debate about whether the members of society wish such a society and no discussion of how to guide the development of the society along these lines. What technology can do and can be developed will be done and developed. The MACIT implicitly sanctions this. According to the myth, these are not issues that have moral import or deserve moral scrutiny. Reality and progress march on, and attempting to stand in the way, slow the march, or evaluate them critically is to misconstrue the future. The result is acceptance of what is developed and how. (p. 6)

D. G. Johnson (✉)
University of Virginia, Charlottesville, VA, USA
e-mail: dgj7p@Virginia.EDU

Published online: 22 May 2014

De George is right to take issue with the MACIT and to use that critique as the starting place—the foundation—of his analysis of the ethical issues arising from adoption and use of various kinds of information technologies. Although I think he is wrong to agree even with the part of the myth specifying that moral language is inappropriate for technology (Johnson 2006), that is not the issue that I will take up in what follows.

De George's analysis raises some larger questions about the relationship between technology and morality, questions that need further examination. He is unambiguous in his claim that people, not technology, are responsible for what people do with technology. He writes: "Those who build, program, run, own, and/or manage the computers or information systems are the only ones who can be held morally responsible for results" (p 30). This claim will be the focus of attention in the analysis that follows. Specifically, I will consider a challenge to this claim from those who argue that in the future we will be confronted with autonomous technologies, e.g., bots and robots, for which no humans can be responsible. Although the challenge is about future technologies, the arguments are important for what they reveal about the relationship between technology and responsibility.

The Responsibility Gap

A major challenge to the claim that human beings and only human beings can be responsible for the behavior of machines (technologies) comes from those who focus on artificial agents that have the capacity to learn as they operate. The term artificial agent refers broadly to computational devices that perform tasks on behalf of humans and do so without immediate, direct human control or intervention. Some artificial agents are software programs, e.g., bots that perform Internet web searches; these programs are purely computational. Other artificial agents are hardware–software combinations, e.g., robots; these combinations have computational decision-making components embedded in their embodied structures. Some argue that because certain artificial agents learn as they operate, those who designed or deployed those agents may not be able to control or even predict what their agents will do. As these agents become increasingly more autonomous, the argument goes, no humans will be responsible for their behavior. Matthias (2004) characterizes this possible, future situation by referring to *a responsibility gap*. He writes:

> … presently there are machines in development or already in use which are able to decide on a course of action and to act without human intervention. The rules by which they act are not fixed during the production process, but can be changed during the operation of the machine, by the machine itself. … Now it can be shown that there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine's actions to be able to assume the responsibility for them.

To support his position, Matthias first describes a number of systems in development or already in use that have the relevant characteristics, and then he describes four different types of learning automata (artificial intelligence systems) showing how in each case, and in different ways, the original designer loses control over the behavior of the device. He argues that the complexities of each lead to complexities in ascribing responsibility for the behavior of the artificial agents. For our purposes, such a situation would seem to constitute a counterexample to De George's claim that humans, not technology, are always responsible for what is done with technology.

Sparrow (2007) makes a similar argument, though he is concerned only with autonomous weapon systems. (AWS). Taking programmers, the commanding officer, and the machine itself as the likely candidates for bearing responsibility for AWS behavior, Sparrow argues that responsibility is not justified for any of them. His explanation of why programmers are not responsible illustrates his acceptance of the responsibility gap. Sparrow writes:

> The possibility that an autonomous system will make choices other than those predicted and encouraged by its programmers is inherent in the claim that it is autonomous. If it has sufficient autonomy that it learns from its experience and surroundings then it may make decisions that reflect these as much, or more than its initial programming. The more the system is autonomous, then the more it has the capacity to make choices other than those predicted or encouraged by its programmers. At some point then, it will no longer be possible [to] hold the programmers/designers responsible for outcomes that they could neither control nor predict. The connection between the programmer/designers, and the results of the system that ground the attribution of responsibility, is broken by the autonomy of the system. (p. 70)

Sparrow uses this responsibility gap as the basis for his argument against the use of AWS; that is, he claims that the use of autonomous weapon systems is unethical precisely because no humans can be responsible for what they do.

Both Matthias and Sparrow seem to believe that the programming techniques at issue will be developed and put to use because of what they can accomplish and despite the fact that we will not be able to control or predict how they will behave. One way to challenge a prediction is to make an alternative prediction and try to show that the alternative is more likely. In this case that would mean claiming that such technologies will not be developed or, if developed, will not be adopted and used. That is not the approach I will take here though my strategy will facilitate the counter prediction. My strategy is to argue that speculations about a responsibility gap misrepresent the situation and are based on false assumptions about technological development and about responsibility.

Responses to the Responsibility Gap

Responses to the specter of a responsibility gap have been wide ranging. The most direct criticism has been to attack an underlying assumption about the fairness of attributions of responsibility. (Santoro et al. 2008) reject the responsibility gap by rejecting what they refer to as the control requirement (CR). According to CR, it is not fair to hold someone responsible for outcomes or behavior that they could not control. Both Mathias' and Sparrow's arguments presume that it is unfair to blame humans for the behavior of machines that they can not control. However, Santoro, et al. deny this. They point out that in other contexts we use "a variety of conceptual frameworks and technical tools … which enable one to deal with problems of responsibility ascription without appealing to (CR)" (p. 310). Their point is that there are situations in which we hold humans responsible for outcomes that they could not control. Strict liability is an obvious example here.

(Nagenborg et al. 2008) argue that *engineers* would be held responsible for the behavior of artificial agents even if they can't control them, on grounds of professional responsibility. For engineers to avoid such responsibility would be a serious breach of professional conduct as it would be in other cases of dangerous and risky products. Thus, Nagenborg et al. also reject CR and argue not just that engineers are fairly held responsible for the behavior of machines they create, but that this responsibility "comes with the territory" of being an engineer.

The Santoro and Nagenborg arguments both refer to responsibility practices in which individuals or corporate entities are held responsible despite the fact that they are not able to control the outcome. In the literature on artificial agents, this has been the direction taken by those who suggest that existing law will prevent or fill the responsibility gap. For example, Asaro (2007, 2012) reviews product liability law, vicarious liability, the law of agency, the concept of diminished responsibility, and the criminal law, showing how these laws might be used in the case of autonomous agents that learn.[1] New technologies often require some modification or extension of existing legal mechanisms, so artificial agents would not be unique if existing law had to be extended in order to address issues of responsibility. The point is that responsibility practices can be developed in which humans are held responsible for artificial agents; there is precedent for these practices and such practices eliminate any supposed responsibility gap.

Another response to the responsibility gap, and more broadly to concerns about increasingly powerful, autonomous decision-making robots, has been to push in the direction of programing artificial agents to be ethical. This is the agenda of the field of machine ethics; the goal is "to create a machine that follows an ideal ethical principle or a set of ethical principles in guiding its behavior" (Anderson, 2011). This endeavor has been taken up by a small number of computer scientists and philosophers who work on the computation of ethical theories and/or software that implements ethical principles in particular contexts such as medical caregiving or warfare (Anderson and Anderson 2011; Allen et al. 2005; Arkin 2009). Arkin (2008, 2009, 2010), for example, designs software for military robots operating on the battlefield. He argues that properly programmed machines may be more ethical than humans in certain combat situations. They will not be encumbered by the drive to self-preservation; they can make use of more information, from more resources, more quickly; and they will not have emotions that may "cloud their judgment" (2009).

Yet another response to the responsibility gap has been to entertain the possibility that artificial agents could themselves be responsible. Hellström (2013), for example, claims that the "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" (p. 105). His argument relies on the idea that robots will be responsive to praise and blame, and hence holding robots responsible will have a deterrent effect. Here, Hellström treats responsiveness to praise and blame as comparable to reinforcement learning. In any case, his argument rests finally on the tendency of humans to assign responsibility to computers and robots rather than something that would justify the attribution of responsibility (p. 105). In other words, his claim is not that robots will *be* responsible for their own behavior but that humans will be inclined to treat robots as if the robots were responsible for their own behavior.

[1] A more extended analysis of the law of agency as it might apply to artificial agents is found in Chopin and White (2012).

Asaro (2012) also considers the possibility of the agents themselves being responsible in his analysis of the criminal law as a mechanism for handling autonomous artificial agent behavior. In the criminal context the question of personhood arises—the personhood of robots—and the efficacy of punishment. Asaro does not take a stance here though he lays the groundwork for robot responsibility by considering the parallel between robots and corporations. Like corporations, robots could be a nonhuman entity that bears responsibility.

As will be discussed later, the possibility of robots that are responsible for their own behavior converges with a stream of analysis suggesting that artificial agents of the future could acquire the status—not of persons but—of moral agents, at least in the sense that they will have a kind of moral standing. Sullins (2006, 2009) takes the extreme position here arguing that autonomous robots could acquire the status of moral agents. Others push for moral standing by focusing on rights against abuse of robots. For example, Whitby (2008) argues that robots should not be abused and Petersen (2007) is concerned about the immorality of "robot servitude." Although neither of these arguments suggests that robots of the future will or could have personhood, both arguments move in the direction of negative rights for robots. Moral standing does not necessarily mean responsibility for one's behavior, but the arguments for moral standing could converge with the work being done to program artificial agents to be ethical. Establishing that artificial agents have moral standing and that they have the capacity to adhere to moral norms (i.e., to behave morally) would provide at least some of the groundwork for holding artificial agents responsible.

Except for the stream of analysis entertaining the possibility that artificial agents could themselves be responsible, these responses to the responsibility gap aim either to change the design of the technology or to change our way of thinking about responsibility, e.g., eliminating the control requirement or adapting existing legal principles. None of the responses challenges the mode of thinking that leads to the idea of a technology for which no humans can be responsible. In the next section, I argue that this idea is based on a misunderstanding of technological development and a misconception of responsibility.

Technological Development and the Future

In the literature on responsibility and artificial agents, the notion of responsibility is underdeveloped. Both those who entertain the possibility of a responsibility gap and those who argue for human responsibility rarely explore what it means to say that an entity is responsible or how it happens that an entity has or does not have responsibility. Sometimes responsibility is seen as a state of being or a quality or characteristic assigned to an entity. Hellström (2013), for example, specifies responsibility as a quality. He writes: "… responsibility will be regarded as a quality we assign to others, in varying degrees based on some, possibly unspecified, norm system" (page 103). Other times, responsibility is taken to mean amenability to punishment; yet other times it is assumed that if an entity has the capacity to make decisions, then it is responsible for those decisions.

To be sure, responsibility is a complex concept. It is one of a cluster of moral concepts that are loosely connected and overlapping. "Responsibility," "responsible," "morally responsible," "blameworthy," "accountable," and "liable" are sometimes used interchangeably and other times used in distinctive ways. Given the variety of terms and their overlapping meanings, it is, perhaps, not surprising that responsibility is a point of contention when it comes to new technologies such as artificial agents.

In addition to the lack of clarity on responsibility, there is a consequent lack of clarity about the relationship between responsibility and technology. Those who are worried about a responsibility gap seem to think that the nature of a technology determines its responsibility conditions. Others in the discourse argue that responsibility (human responsibility) requires that technologies be designed and built so as to facilitate, or even ensure, human responsibility. For example, Cummings (2004, 2006) takes this approach in her research on interfaces for communications between robots and humans.

So what is the relationship between technology and responsibility? In order to answer this question, some attention must be given to the futuristic nature of the discourse on artificial agents and the processes by which new technologies are produced.

Speculation, Predictions, and Visions

The discourse on responsibility and artificial agents described above is largely a discourse about the future. Those who foresee a responsibility gap do not claim that current agents have the kind of autonomy that means no human responsibility; they claim only that agents of the future will or may have that kind of autonomy. Importantly, claims about the capabilities of future technologies are, by their very nature, speculative or predictive. Speculation and prediction about the future generally involve taking current trends and patterns, and then extrapolating out in time from what happened in the past and is happening now to what is likely to happen in the future. For example, one might use the history of the last several decades of robot development as the basis for predicting that robots will become more fully autonomous at some

point in the future. That is, the historical record provides evidence of progressive development in the autonomy and learning capabilities of artificial agents, and this would seem to justify the assumption that future development will continue on the same trajectory.

Speculation can be contrasted with prediction and with vision. Prediction and speculation overlap in the sense that both forecast the future, but speculation is more tentative than prediction. In principle predictions are amenable to truth testing; that is, we will know in the future whether a prediction was true or false. Speculations, on the other hand, are not subject to truth testing. They are intended to be looser, more tentative, and open to alternative possibilities. As well, claims about the future can, individually or as a set, constitute a vision. Visions provide a picture of a future world, a state of affairs that could be realized. Visions direct actors to engage in activities that will bring about that future because the possible future is considered desirable or inevitable or both.

Various streams of analysis in the discourse described above seem to merge into a vision. Researchers and scholars are working on projects that if realized could converge on a future in which artificial agents are able to learn in ways that humans do not understand, are increasingly more autonomous, adhere to moral rules, have moral standing, and (not mentioned above) are made to look like humans (are humanoid). Some visions of the future go even beyond this. Consider, for example, the description of the book *Robot Futures* by Illah Reza Nourbakhsh (2013) provided by amazon.com:

> The ambition of modern robotics goes beyond copying humans, beyond the effort to make walking, talking androids that are indistinguishable from people. Future robots will have superhuman abilities in both the physical and digital realms. They will be embedded in our physical spaces, with the ability to go where we cannot, and will have minds of their own, thanks to artificial intelligence. They will be fully connected to the digital world, far better at carrying out online tasks than we are. In *Robot Futures*, the roboticist Illah Reza Nourbakhsh considers how we will share our world with these creatures, and how our society could change as it incorporates a race of stronger, smarter beings. Nourbakhsh imagines a future that includes adbots offering interactive custom messaging; robotic flying toys that operate by means of "gaze tracking"; robot-enabled multimodal, multicontinental telepresence; and even a way that nanorobots could allow us to assume different physical forms.[2]

[2] (Accessed at http://www.amazon.com/Robot-Futures-Illah-Reza-Nourbakhsh/dp/0262018624/ref=sr_1_1?s=books&ie=UTF8&qid=1375020104&sr=1-1&keywords=future+of+robots, July 28, 2013).

Visions like Nourbakhsh's might be dismissed as fanciful; however, in the context of technological development, visions, even fanciful ones, have the power to influence the future. That is, visions promote research agendas and investments of time and money in making the vision a reality.

Speculation, predictions, and especially visions, have rhetorical power; they can be used as a form of persuasion, to enroll others into activities that help make a particular future a reality. They can also be used to lay groundwork making us comfortable with a situation that might occur in the future. Visions can be dangerous, as well, insofar as they draw attention away from other possibilities and other possible agendas for research and development. In this respect, discussion of the responsibility gap is of concern insofar as it persuades us to accept and be comfortable with the idea of technologies for which no human can be responsible. Discussions about the responsibility gap (as if it were inevitable) draw our attention away from the possibility of designing technologies so as to ensure human responsibility for what they do.

Technological Development

Whether speculation, prediction or vision, some claims about the future are more plausible than others because of their assumptions. In the case of future technologies, extrapolation to the future typically relies on assumptions about what will become technologically feasible. That is, projections about the future presume that what is not now technologically feasible will become feasible in the future. What is often missed is that technological feasibility itself is dependent on human activity; for something to become technically feasible in the future, researchers must continue to do their work and this, in turn, often means that funders must continue to support the research.

What is striking about the idea of artificial agents for which no humans can be responsible is that the very idea relies on a narrow and deficient view of technological development. Those who are concerned about a responsibility gap seem to believe that technological development proceeds in a fashion, wherein one technical accomplishment builds on a prior technical accomplishment and the new accomplishment lays the foundation for the next technical breakthrough. Authors such as Matthias and Sparrow seem to believe that the sequence of steps involved in technological development has a logic of its own, i.e., determined entirely by nature or the nature of a particular technology. They seem to believe that researchers and engineers simply follow the logic to its inevitable conclusion. In the case of artificial agents, the inevitable conclusion is thought to be agents that are fully autonomous, presumably useful, and yet not understandable to humans.

The last several decades of research in the field of science and technology studies (STS) have replaced this view of technological development with a view in which it is seen as non-linear, multidirectional, and contingent (Bijker et. al. 1987; MacKenzie and Wajcman 1996). In hindsight someone may provide a linear account of a technology's development and make it seem as if the latest design was the natural, inevitable outcome of prior designs. However, such accounts are over-simplifications of a reality that is much more complicated and messy than the linear narrative suggests.

On STS accounts, technological development involves many different actors with interests that push development in a variety of directions. The many actors—scientists and engineers, funding agencies, regulatory bodies, manufacturers, the media, the public, and others—affect the direction of development. The actors negotiate to have their interests served in the way the new technology is designed. The negotiations ultimately result in coalescence around a particular design and meaning for a new technology (Johnson 2005). This is not to say that technological feasibility and past research and invention are not important to future developments. They are. The point is that there is a lot more than technological feasibility involved in shaping future technologies and, most importantly, the outcomes of research and development are contingent, not inevitable. The economic environment, regulatory decisions, historical events, public attitudes, media presentations, and much more affects what is developed, adopted, and used.

In the case of artificial agents and autonomy, multiple directions of development can be observed even when it comes to autonomous agents.[3] For example, a recent Department of Defense publication suggests that the future development of military technology will be focused less on robot autonomy and more on robots supporting human decision making (U.S. Department of Defense 2012). Especially in the case of the military, such statements usually mean investment in the new direction and this, in turn, means researchers and developers turning their attention to the new direction. Other evidence of this alternative trajectory of development for artificial agents is found in (Johnson et al. 2011): "We no longer look at the primary problem of the research community as simply trying to make agents more independent through their autonomy. Rather, in addition, we strive to make them more capable of sophisticated interdependent joint activity with people" (p. 189).

Thus, even in the case of artificial agents, development is multidirectional and it is far from clear what agents of the future will be like, that is, how they will operate, how they will be used, and what role they will have in particular contexts. The social nature of technological development makes the task of prediction and speculation especially challenging. We have to consider more than the technical trajectory; we have to consider an array of other activities that are going on, the actors who are involved, the ideas that are being explored, the investments that are being made, and the public's reactions, to mention a few. We cannot be sure whether or how nature will cooperate and we cannot be sure exactly which actors and which ideas will be most influential in shaping future developments. For example, we do not know whether or how autonomous car developers such as Google will succeed in programming autonomous cars to handle crash situations and we do not know whether the developers will be able to convince regulatory agencies and the public that autonomous cars are safe enough.[4] We don't know whether autonomous cars will be used in the future, and if they are used, we can not know what that will be like because the design of the cars will likely be affected by many considerations including how to convince regulatory agencies and the public that the vehicles will be safe and reliable.

In assuming that artificial agent development will proceed in a linear trajectory with only technical factors coming into play, those concerned about the responsibility gap fail to see that the development of artificial agent technologies will be affected by a variety of human actors and actor groups with varying interests, values, and capacities to influence development. More autonomous technologies may well be developed in the future and a responsibility gap may occur, but, if the gap occurs, this will be the result of human choices and not the inevitable outcome of the kinds of technologies currently in development.

Technology and Responsibility Practices

Given that producing a new technology involves many human actors making decisions and getting others to accept those decisions, in order to imagine a future time at which there will be artificial agents for which no humans are responsible, we have to imagine that the human actors involved would decide to create, release, and accept technologies that are incomprehensible and out of the control of humans. In addition, we have to imagine that the humans

[3] Elsewhere I have explored the varying conceptions of autonomy that are being used in this discourse; see M. Noorman and D.G. Johnson, "Negotiating Autonomy and Responsibility in Military Robots", *Ethics and Information Technology*, forthcoming.

[4] Google has succeeded in convincing several municipalities to allow Google's so-called autonomous cars to operate in their areas but these cars are not unmanned.

involved (especially consumers, users, and the public) would accept an arrangement in which no humans would be considered responsible for these technologies. Importantly, these two steps are separable. Putting (or allowing to be put) into operation incomprehensible and uncontrollable artificial agents does not necessitate accepting an arrangement in which no humans are responsible for the behavior of those agents.[5] Santoro et. al. recognized this in claiming the control requirement is not essential to responsibility.

Having separated out the decision to accept incomprehensible and uncontrollable technologies from the decision to accept an arrangement in which no one is responsible, we can now focus further on responsibility arrangements. As mentioned earlier, the notion of responsibility is both complex and underdeveloped in the discourse on artificial agents. A better understanding of responsibility will show further that the idea of technologies for which no human can be responsible is misguided.

In the context of artificial agents, the kind of responsibility that seems to be at issue is accountability. That is, the responsibility gap raises the question whether there can be technologies for which no human can be accountable. When a technology behaves in ways that we did not expect (and especially when the unexpected behavior results in accidents or mishaps), we often want to know why and who is to blame. We may want something to be done to ensure that the unexpected behavior does not occur again or we may simply want to understand better how the technology works so that we can change our expectations. In order to know who to blame or what might be done to change the situation, we need to know who is accountable.

Accountability-responsibility is embedded in relationships that involve norms and expectations. The relationship may be general, e.g., a responsibility to all human beings, or it may be specific as in the case of an employee having a responsibility to perform particular tasks for an employer. According to Bovens (2007), accountability "is a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgment, and the actor may face consequences." In accountability relationships those who are accountable believe they have an obligation to a forum, e.g., a community, the public, a particular individual or group of individuals. Members of the forum believe that they are owed an explanation; they expect that those who are accountable will answer (provide an account) when they fail to adhere to appropriate norms, i.e., fail to live up to expectations. In these relationships, norms and expectations are generally shared.

Norms and expectations in accountability relationships are constituted formally and informally. They are informally transmitted in culture and they can be more formally transmitted for particular contexts, for example, in a job description, in a professional code of conduct, or in a user manual. While Nagenborg et al. note that engineers are held accountable for the technologies they develop, they describe the relationship between engineers and the public. Formally, this is specified in the law with engineers legally liable for technologies they design; this is true at least for licensed engineers who sign off on drawings. Informally, the public, as well as clients and consumers, expect engineers to account for what they do in part because engineers have shaped these expectations by promulgating codes of professional conduct, and through other activities that shape public attitudes and promote engineering.

However, since modern technologies involve "many hands" both in their production and in their use, many actors may be accountable for different aspects of the operation of a technology. This is most evident when accidents occur. The cause of the accident has to be traced back to the relevant actor/s; the cause may be in any number of places: Was the design adequate? Did the manufactured parts meet specifications? Did the instructions adequately explain how to use the technology? Did the users treat the technology as instructed? Each of the actors or actor groups is accountable for their contribution to the production of the technology and each may be asked to account if something unexpected happens.

Recognizing that responsibility is embedded in relationships adds further support to the idea that the nature of a particular technology does not necessitate a particular responsibility arrangement. Accountability relationships are not dictated by nature or anything else. The nature of a technology is relevant to the responsibility arrangements, but responsibility arrangements are socially constituted through the norms and expectations of particular activities and contexts.

Hence, when it comes to responsibility-accountability for artificial agents of the future, the possibilities are open. People of the future might accept no human responsibility; they might come to expect robots to explain their behavior specifying why they did what they did; they might hold robot manufacturers accountable or they might hold multiple parties accountability according to their particular contributions to robot behavior. In other words, the responsibility arrangements for particular technologies of the future are contingent. They will be negotiated and worded out as the technology is being developed, tested, put into operation, and used.

[5] Although the idea will not be taken up here, it is worth noting that the notion of an incomprehensible and uncontrollable technology needs to be unpacked for many current technologies are incomprehensible and uncontrollable to some but not to others.

There are good reasons for staying with human responsibility, namely to keep the pressure on developers to ensure the safety and reliability of such devices. However, to avoid speculation about what humans of the future will accept, it is enough to say that although the idea of a technology for which no human can be responsible is not an incoherent concept, it is misguided insofar as it implies that because of the nature of technology, human responsibility will not be possible.

Conclusion

So, do autonomous artificial agents of the future constitute a counterexample to the claim – made by De George – that people, not technology, are responsible for technological outcomes? Is it possible that artificial agents of the future will have learning capabilities and the capacity for autonomous behavior such that no humans can be responsible for them? Although we do not and cannot know what future technologies will be like, the preceding analysis makes clear that whether or not there will ever be a responsibility gap depends on human choices not technological complexity. A responsibility gap will not arise merely from the technological complexity of artificial agents. Artificial agents can be designed so that no human can understand or control what they do or they can be designed so that they are transparent and well within human control or they can be designed so that certain aspects or levels of the machine behavior are in human control and others are not. Which way they are designed depends on the humans involved in their development and acceptance.

In the past people have chosen technologies that have some degree of risk though we have also set up mechanisms to pressure those who make and use these technologies to operate them safely and to take responsibility when something goes wrong and the fault can be traced back to them. The future may be different, but it seems there are good reasons why we might resist any future in which no humans are responsible for technologies that have a powerful role in our lives.

None of this is to say that all is well and there is no need to worry. Recognizing the contingency of technological development entails recognizing that in the process of development, decisions by humans could lead to an arrangement in which humans accept robots knowing that no humans understand how they arrive at their decisions and how they will behave in certain circumstances. However, if things go this way, it will not be a natural evolution of technological development. Rather it will be because in the negotiations about the technology, certain actors pushed in that direction, were able to enroll others in their way of thinking, and together they won the day in terms of design *and* responsibility practices.

Acknowledgments Research for this paper was 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 author and do not necessarily reflect the views of the National Science Foundation. This article has been greatly improved from comments on an earlier version from Norm Bowie and Keith Miller.

References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: top–down, bottom–up, and hybrid approaches. *Ethics and Information Technology, 7*(3), 149–155.

Anderson, S. L. (2011). Machine metaethics. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 21–27). New York: Cambridge University Press.

Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine ethics*. New York: Cambridge University Press.

Arkin, R. C. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture part I: Motivation and philosophy. In *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on* (pp. 121-128). IEEE.

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

Arkin, R. C. (2010). The case for ethical autonomy in unmanned systems. *Journal of Military Ethics, 9*(4), 332–341.

Asaro, P. M. (2012). 11 A body to kick, but still no soul to damn: Legal perspectives on robotics. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics*. Cambridge: MIT Press.

Asaro, P. (2007). Robots and responsibility from a legal perspective. *Proceedings of the IEEE*.

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

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

Cummings, M. L. (2006). Automation and accountability in decision support system interface design. *Journal of Technology Studies, 32*(1), 23–31.

De George, R. T. (2003). *The ethics of information technology and business*. Malden: Blackwell Publishing.

Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and Information Technology, 12*(2), 99–107.

Johnson, D. G. (2005). The social construction of technology. In C. Mitcham (Ed.), *The encyclopedia of science, technology, and ethics*. Farmington Hills: Gale Group Publishing.

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

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, B., & Sierhuis, M. (2011). The fundamental principle of coactive design: Interdependence must shape autonomy. *Coordination, organizations, institutions, and norms in agent systems VI*. Heidelberg: Springer.

MacKenzie, D., & Wajcman, J. (1996). *The social shaping of technology* (2nd ed.). Buckingham: Open University Press.

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

Nagenborg, M., Capurro, R., Weber, J., & Pingel, C. (2008). Ethical regulations on robotics in Europe. *AI & Society, 22*(3), 349–366.

Nourbakhsh, I. R. (2013). *Robot futures*. MIT Press.

Petersen, S. (2007). The ethics of robot servitude. *Journal of Experimental & Theoretical Artificial Intelligence, 19*(1), 43–54.

Santoro, M., Marino, D., & Tamburrini, G. (2008). Learning robots interacting with humans: from epistemic risk to responsibility. *AI & Society, 22*(3), 301–314.

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

Sullins, John P. (2006). When is a robot a moral agent? *International Review of Information Ethics, 6*, 23–30.

Sullins, John P. (2009). Artificial moral agency in technoethics. In R. Luppicini & R. Adell (Eds.), *Handbook of research on technoethics* (pp. 205–221). New York: IGI Global.

U.S. Department of Defense (2012). Task force report: The role of autonomy in DoD systems. http://www.fas.org/irp/agency/dod/dsb/autonomy.pdf. Accessed July 16, 2013.

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.