Ethics and Information Technology (2006) 8:195–204
DOI 10.1007/s10676-006-9111-5

Computer systems: Moral entities but not moral agents

Deborah G. Johnson
Department of Science, Technology, and Society, University of Virginia, 351 McCormick Road, Charlottesville, VA 22904-4744, USA
E-mail: dgj7p@virginia.edu

Abstract. After discussing the distinction between artifacts and natural entities, and the distinction between artifacts and technology, the conditions of the traditional account of moral agency are identified. While computer system behavior meets four of the five conditions, it does not and cannot meet a key condition. Computer systems do not have mental states, and even if they could be construed as having mental states, they do not have intendings to act, which arise from an agent's freedom. On the other hand, computer systems have intentionality, and because of this, they should not be dismissed from the realm of morality in the same way that natural objects are dismissed. Natural objects behave from necessity; computer systems and other artifacts behave from necessity after they are created and deployed, but, unlike natural objects, they are intentionally created and deployed. Failure to recognize the intentionality of computer systems and their connection to human intentionality and action hides the moral character of computer systems. Computer systems are components in human moral action. When humans act with artifacts, their actions are constituted by the intentionality and efficacy of the artifact which, in turn, has been constituted by the intentionality and efficacy of the artifact designer. All three components – artifact designer, artifact, and artifact user – are at work when there is an action and all three should be the focus of moral evaluation.

Key words: action theory, artifact, artificial moral agent, intentionality, moral agent, technology

Introduction

In this paper I will argue that computer systems are moral entities, but not, alone, moral agents. In making this argument I will navigate through a complex set of issues much debated by scholars of artificial intelligence, cognitive science, and computer ethics. My claim is that those who argue for the moral agency (or potential moral agency) of computers are right in recognizing the moral importance of computers, but they go wrong in viewing computer systems as independent, autonomous moral agents. Computer systems have meaning and significance only in relation to human beings; they are components in socio-technical systems. What computer systems are and what they do is intertwined with the social practices and systems of meaning of human beings. Those who argue for the moral agency (or potential moral agency) of computer systems also go wrong insofar as they overemphasize the distinctiveness of computers. Computer systems are distinctive, but they are a distinctive form of technology and have a good deal in common with other types of technology.

On the other hand, those who claim that computer systems are not (and can never be) moral agents, also, go wrong when they claim that computer systems are outside the domain of morality. To suppose that morality applies only to the human beings who use computer systems is a mistake.

The debate seems to be framed in a way that locks the interlocutors into claiming either that computers are moral agents or that computers are not moral. Yet, to deny that computer systems are moral agents is not the same as denying that computers have moral importance or moral character; and to claim that computer systems are moral is not necessarily the same as claiming that they are moral agents. The interlocutors neglect important territory when the debate is framed in this way. In arguing that computer systems are moral entities but are not moral agents, I hope to reframe the discussion of the moral character of computers.

I should add here that the debate to which I refer is embedded in a patchwork of literature on a variety of topics. Since all agree that computers are currently quite primitive in relation to what they are likely to be

in the future, the debate tends to focus on issues surrounding the potential capabilities of computer systems and a set of related and dependent issues. These issues include whether the agenda of artificial intelligence is coherent; whether, moral agency aside, it makes sense to attribute moral responsibility to computers; whether computers can reason morally or behave in accordance with moral principles; and whether computers (with certain kinds of intelligence) might come to have the status of persons and, thereby, the right not to be turned off. The scholars who come the closest to claiming moral agency for computers are probably those who use the term "artificial moral agent" (AMA), though the term hedges whether computers are moral agents in a strong sense of the term, comparable to human moral agents, or agents in the weaker sense in which a person or machine might perform a task for a person and the behavior has moral consequences.[1,2]

Natural and human-made entities/artifacts

The analysis and argument that I will present relies on two fundamental distinctions, the distinction between natural phenomena or natural entities *and* human-made entities, and the distinction between artifacts *and* technology. Both of these distinctions are problematic in the sense that when pressed, the line separating the two sides of the distinction can be blurred. Nevertheless, these distinctions are foundational. A rejection or re-definition of these distinctions obfuscates and undermines the meaning and significance of claims about morality, technology, and computing.

The very idea of technology is the idea of things that are human-made. To be sure, definitions of technology are contentious, so I hope to go to the heart of the notion and avoid much of the debate. The association of the term "technology" with human-made things has a long history dating back to Aristotle.[3] Moreover, making technology has been understood to be an important aspect of being human. In "The Question Concerning Technology" Heidegger writes:

> "For to posit ends and procure and utilize the means to them is a human activity. The manufacture and utilization of equipment, tools, and machines, the manufactured and used things all belong to what technology is. The whole complex of these contrivances is technology. Technology itself is a contrivance – in Latin, an instrumentum."[4]

More recently and consistent with Heidegger, Pitt gives an account of technology as "humanity at work."[5]

While the distinction between natural and human-made entities is foundational, I concede that the distinction can be confounded. When a tribesman picks up a stick and throws it at an animal, using the stick as a spear to bring the animal down, a natural object – an object appearing in nature independent of human behavior – has become a tool. It has become a means for a human end. Here a stick is both a natural object and a technology.

Another way the distinction can be challenged is by consideration of new biotechnologies such as genetically modified foods or pharmaceuticals. These technologies appear to be combinations of nature and technology, combinations that make it difficult to disentangle and draw a line between the natural and human-made parts. These new technologies are products of human contrivance, though the human contrivance is at the molecular level and this makes the outcome or product appear natural in itself. Interestingly, the only difference between biotechnology and other forms of technology – computers, nuclear missiles, toasters, televisions – is the kind of manipulation or the level at which the manipulation

[1] Those who use the term "artificial moral agent" include: L. Floridi and J. Sanders. On the Morality of Artificial Agents. *Minds and Machines*, 14(3): 349–379, 2004; B.C. Stahl. Information, Ethics, and Computers: The Problem of Autonomous Moral Agents. *Minds and Machines*, 14: 67–83, 2004; and C. Allen, G. Varner and J. Zinser. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12: 251–261, 2000.

[2] For an account of computers as surrogate agents, see D.G. Johnson and T.M. Powers. Computers as Surrogate Agents. In J. van den Hoven and J. Weckert, editors, *Moral Philosophy and Information Technology*. Cambridge University Press, 2006.

[3] In the *Nicomachean Ethics*, Aristotle writes, "Every craft is concerned with coming to be; and the exercise of the craft is the study of how something that admits of being and not being comes to be, something whose origin is in the producer and not in the product. For a craft is not concerned with things that are or come to be by necessity; or with things that are by nature, since these have their origin in themselves." (6.32). [Translation from Terence Irwin, Indianapolis, Hackett, 1985.]

[4] From M. Heidegger, *The Question Concerning Technology and Other Essays*. Translated and with an Introduction by W. Lovitt. Harper & Row, New York, 1977.

[5] J. Pitt. *Thinking About Technology: Foundations of the Philosophy of Technology*. Seven Bridges Press, New York, 2000.

of nature takes place. In some sense, the action of the tribesman picking up the stick and using it as a spear and the action of the bioengineer manipulating cells to make a new organism are of the same kind; both manipulate nature to achieve a human end. The difference in the behavior is in the different types of components that are manipulated.

Yet another way the distinction between natural and human-made entities can be pressed has to do with the extent to which the environment has been affected by human behavior. Environmental historians are now pointing to the huge impact that human behavior has had on the earth over the course of thousands of years of human history. They point out that we can no longer think of our environment as "natural."[6] In this way, distinguishing nature from what is human-made is not always easy.

Nevertheless, while all of these challenges can be made to the distinction between natural and human-made, they do not indicate that the distinction is incoherent or untenable. Rather, the challenges indicate that the distinction between natural and human-made is useful and allows us to understand something important. Eliminating this distinction would make it impossible for us to distinguish the effects of human behavior on, or the human contribution to, the world that is. Eliminating this distinction would make it difficult, if not impossible, for humans to comprehend the implications of their normative choices about the future. There would be no point in asking what sort of world we want to make, whether we (humans) should do something to slow global warming, or slow the use of fossil fuel or prevent the destruction of ecosystems. These choices only make sense when we recognize a distinction between the effects of human behavior and something independent of human behavior – nature.

The second distinction at the core of my analysis is the distinction between artifacts and technology. A common way of thinking about technology – perhaps the layperson's way – is to think that it is physical or material objects. I will use the term *artifact* to refer to the physical object. Philosophers of technology and recent literature from the field of science and technology studies (STS) have pointed to the misleading nature of this view of technology. Technology is a combination of artifacts, social practices, social relationships, and systems of knowledge. These combinations are sometimes referred to as *socio-technical ensembles*[7] or *socio-technical systems*[8] or *networks*.[9] Artifacts (the products of human contrivance) do not exist without systems of knowledge, social practices, and human relationships. Artifacts are made, adopted, distributed, used and have meaning only in the context of human social activity. Indeed, while we intuitively may think that artifacts are concrete and "hard," and social activity is abstract and "soft," the opposite is more accurate. Artifacts are abstractions from reality. To delineate an artifact – that is, to identify it as an entity – we must perform a mental act of separating the object from its context. The mental act extracts the artifact from the social activities that give it meaning and function. Artifacts come into being through social activity, are distributed and used by human beings as part of social activity, and have meaning only in particular contexts in which they are recognized and used. When we conceptually separate an artifact from the contexts in which it was produced and used, we push the socio-technical system of which it is a part out of sight.

So it is with computers and computer systems. They are as much a part of social practices as are automobiles, toasters, and playpens. Computer systems are not naturally occurring phenomena; they could not and would not exist were it not for complex systems of knowledge and complex social, political, cultural institutions; computer systems are produced, distributed, and used by people engaged in social practices and meaningful pursuits. This is as true of current computer systems as it will be of future computer systems. No matter how independently, automatically, and interactively computer systems of the future behave, they will be the products (direct or indirect) of human behavior, human social institutions, and human decision.

Notice that the terms "computer" and "computer system" are sometimes used to refer to the artifact, other times to the socio-technical system. While we can think of computers as artifacts, to do so is to engage in the thought experiment alluded to above; it is to engage in the act of mentally separating

[6] See, for example, B.R. Allenby. Engineering Ethics for an Anthropogenic Planet. In *Emerging Technologies and Ethical Issues in Engineering*, pp. 7–28. National Academies Press, Washington D.C., 2004.

[7] W.E. Bijker. Sociohistorical Technology Studies. In S. Jasanoff, G.E. Markle, J.C. Petersen and T. Pinch, editors, *Handbook of Science and Technology Studies*, pp. 229–256. Sage, London, 1994.

[8] T.P. Hughes. Technological Momentum. In L. Marx and M.R. Smith, editors, *Does Technology Drive History? The Dilemma of Technological Determinism*. The MIT Press, Cambridge, 1994.

[9] J. Law. Technology and Heterogeneous Engineering: The Case of Portuguese Expansion. In W.E. Bijker, T.P. Hughes and T. Pinch, editors, *The Social Construction of Technological Systems*. MIT Press, Cambridge, 1987.

computers from the social arrangements of which they are a part, the activities that produce them, and the cultural notions that give them meaning. Computer systems always operate in particular places at particular times in relation to particular users, institutions, and social purposes.

The separation of computers from the social context in which they are used can be misleading. My point here is not unrelated to the point that Floridi and Saunders make about levels of abstraction.[10] They seem implicitly to concede the abstractness of the term "computer" and would have us pay attention to how we conceptualize computer activities, that is, at what level of abstraction we are focused. While Floridi and Saunders suggest that any level of abstraction may be useful for certain purposes, my argument is, in effect, that certain levels of abstraction are not relevant to the debate about the moral agency of computers, in particular, those levels of abstraction that separate machine behavior from the social practices of which it is a part and the humans who design and use it. My reasons for making this claim will become clear in the next two sections of the paper.

In what follows I will use "artifact" to refer to the material object and "technology" to refer to the socio-technical system. This distinction is consistent with, albeit different from, the distinction between nature and technology. Artifacts are products of human contrivance; they are also components in socio-technical systems that are complexes – ensembles, networks – of human activity and artifacts.

Morality and moral agency

The notions of "moral agency" and "action" and the very idea of morality are deeply rooted in western traditions of moral philosophy. Historically human beings have been understood to be different from all other living entities because they are free and have the capacity to act from their freedom. Human beings can reason about and then choose how they behave. Perhaps the best known and most salient expression of this conception of moral agency is provided by Kant. However, the idea that humans act (as opposed to behaving from necessity) is presumed by almost all moral theories. Even utilitarianism presumes that human beings are capable of choosing how to behave. Utilitarians beseech individuals to use a utilitarian principle in choosing how to act; they encourage the development of social systems of rewards and punishments to encourage individuals to choose certain types of actions over others. In presuming that humans have choice, utilitarianism presumes that humans are free.

My aim is not, however, to demonstrate the role of this conception of moral agency in moral philosophy, but rather to use it. I will quickly lay out what I take to be essential aspects of the concepts of moral agency and action in moral philosophy, and then use these notions to think through computer behavior. I will borrow here from Johnson and Powers' account of the key elements of the standard account.[11] These elements are implicit in both traditional and contemporary accounts of moral agency and action.

The idea that an individual is primarily responsible for his or her intended, voluntary behavior is at the core of most accounts of moral agency. Individuals are not held responsible for behavior they did not intend or for the consequences of intentional behavior that they could not foresee. Intentional behavior has a complex of causality that is different from that of non-intentional or involuntary behavior. Voluntary, intended behavior (action) is understood to be outward behavior that is caused by a particular kind of internal states, namely, mental states. The internal, mental states cause outward behavior, and because of this, the behavior is amenable to a reason explanation as well as a causal explanation. All behavior (human and non-human; voluntary and involuntary) can be explained by its causes, but only action can be explained by a set of internal mental states. We explain why an agent acted by referring to their beliefs, desires, and other intentional states.

Contemporary action theory typically specifies that for human behavior to be considered action (and as such appropriate for moral evaluation), it must meet the following conditions. First, there is an agent with an internal state. The internal state consists of desires, beliefs, and other intentional states. These are mental states, and one of these is, necessarily, an intending to act. Together, the intentional states (e.g., a belief that a certain act is possible, a desire to act, plus an intending to act) constitute a reason for acting. Second, there is an outward, embodied event – the agent does something, moves his or her body in some way. Third, the internal state is the cause of the outward event; that is, the movement of the body is rationally directed at some state of the world. Fourth, the outward behavior (the result of rational direction) has an outward effect. Fifth and finally, the effect has to be on a patient – a recipient of an action, a recipient that can be harmed or helped.

[10] Floridi and Saunders (2004).

[11] D.G. Johnson and T.M. Powers. The Moral Agency of Technology. Unpublished manuscript, 2005.

This set of conditions can be used as a backdrop, a standard against which the moral agency of computer systems can be considered. Those who claim that computer systems can be moral agents have, in relation to this set of conditions, two possible moves. Either they can attack the account and show what is wrong with it, and provide an alternative account of moral agency *or* they can accept the account and show that computer systems meet the conditions. Indeed, much of the scholarship on this issue can be classified as taking one or the other of these approaches.[12]

When the traditional account is used as the standard, computer system behavior seems to meet conditions 2–5 with little difficulty; that is, plausible arguments can be made to that effect. With regard to the second condition, morality has traditionally focused on embodied human behavior as the unit of analysis appropriate for moral evaluation, and computer system behavior is embodied. As computer systems operate, changes in their internal states produce such outward behavior as a reconfiguration of pixels on a screen, audible sounds, change in other machines, and so on. Moreover, the outward, embodied behavior of a computer system is the result of internal changes in the states of the computer, and these internal states cause, and are rationally directed at producing, the outward behavior. Thus, the third condition is met.

Admittedly, the distinction between internal and external ("outward") can be challenged (and may not hold up to certain challenges). Since all of the states of a computer system are embodied, what is the difference between a so-called internal state and a so-called external or outward state? This complication also arises in the case of human behavior. The internal states of humans can be thought of as brain states and in this respect they are also embodied. What makes brain states internal and states of the arms and legs of a person external? The distinction between internal states and outward behavior is rooted in the mind–body tradition so that using the language of internal–external may well beg the question whether a non-human entity can be a moral agent. However, in the case of computer systems, the distinction is not problematic because we distinguish internal and external events in computer systems in roughly the same way we do in humans. Thus, conditions 2 and 3 are no more problematic for the moral agency of computer systems than for humans.

The outward, embodied events that are caused by the internal workings of a computer system can have effects beyond the computer system (condition 4) and these effects can be on moral patients (condition 5). In other words, as with human behavior, when computer systems behave, their behavior has effects on other parts of the embodied world, and those embodied effects can harm or help moral patients. The effect may be morally neutral such as when a computer system produces a moderate change in the temperature in a room or performs a mathematical calculation. However, computer behavior can also produce effects that harm or help a moral patient, e.g., the image produced on a screen is offensive, a signal turns off a life support machine, a virus is delivered and implanted in an individual's computer.

In short, computer behavior meets conditions 2–5 as follows: when computers behave, there is an outward, embodied event; an internal state is the cause of the outward event; the embodied event can have an outward effect; and the effect can be on a moral patient.

The first element of the traditional account is the kingpin for the debate over the moral agency of computers. According to the traditional account of moral agency, for there to be an action (behavior arising from moral agency), the cause of the outward, embodied event must be the internal states of the agent, *and* – the presumption has always been that – these internal states are mental states. Moreover, the traditional account specifies that one of the mental states must be an intending to act. While most of the attention on this issue has focused on the requirement that the internal states be mental states, the intending to act is critically important because the intending to act arises from the agent's freedom.

Action is an exercise of freedom and freedom is what makes morality possible. Moral responsibility does not make sense when behavior is involuntary, e.g., a reflex, a sneeze or other bodily reaction. Of course, this notion of human agency and action is historically rooted in the Cartesian doctrine of mechanism. The Cartesian idea is that animals, machines, and natural events are determined by natural forces; their behavior is the result of necessity. Causal explanations of the behavior of mechanistic entities and events are given in terms of laws of nature. Consequently, neither animals nor machines have the freedom or intentionality that would make them morally responsible or appropriate subjects of moral appraisal. Neither the behavior of nature nor the behavior of machines is amenable to reason explanations and moral agency is not possible when a reason–explanation is not possible.

[12] For example, Fetzer explores whether states of computers could be construed as mental states since they have semantics (J.H. Fetzer. *Computers and Cognition: Why Minds Are Not Machines.* Kluwer Academic Press, 2001); and Stahl explores the same issue using their informational aspect as the basis for exploring whether the states of computers could qualify (Stahl 2004).

Again it is important to note that the requirement is not just that the internal states of a moral agent are mental states; one of the mental states must be an intending to act. The intending to act is the locus of freedom; it explains how two agents with the same desires and beliefs may behave differently. Suppose John has a set of beliefs and desires about Mary; he picks up a gun, aims it at Mary, and pulls the trigger. He has acted. A causal explanation of what happened might include John's pulling the trigger and the behavior of the gun and bullet; a reason explanation would refer to the desires and beliefs and intending that explain why John pulled the trigger. At the same time, Jack could have desires and beliefs identical to those of John, but not act as John acts. Jack may also believe that Mary is about to do something reprehensible, may desire her to stop, may see a gun at hand and yet Jack's beliefs and desires are not accompanied by the intending to stop her. It is the intending to act together with the complex of beliefs and desires that leads to action. Why John forms an intending to act and Jack does not is connected to their freedom. John's intending to act comes from his freedom; he chooses to pick up the gun and pull the trigger. Admittedly, the non-deterministic character of human behavior makes it somewhat mysterious, but it is only because of this mysterious, non-deterministic aspect of moral agency that morality and accountability are coherent.

Cognitive scientists and computer ethicists often acknowledge this requirement of moral agency. Indeed, they can argue that the non-deterministic aspect of moral agency opens the door to the possibility of the moral agency of computer systems since some computer systems are, or in the future will be, non-deterministic. To put the point another way, if computer systems are non-deterministic, then they can be thought of as having something like a noumenal realm. When computers are programmed to learn, they learn to behave in ways that are well beyond the comprehension of their programmers and well beyond what is given to them as input. Neural networks are proffered as examples of non-deterministic computer systems. At least some computer behavior may be said to be constituted by a mixture of deterministic and non-deterministic elements, as is human behavior.

The problem with this approach is that while some computer systems may be non-deterministic and, therefore, "free" in some sense, they are not free in the same way humans are. Perhaps it is more accurate to say that we have no way of knowing whether computers are or will be non-deterministic in same way that humans are non-deterministic. We have no way of knowing whether the noumenal realm of computer systems is or will be anything like the noumenal realm of humans. What we do know is that both are embodied in different ways. Thus, we have no way of knowing whether the non-deterministic character of human behavior and the non-deterministic behavior of computer systems are or will be alike in the morally relevant (and admittedly mysterious) way.

Of course, we can think and speak "as if" the internal states of a computer are comparable to the mental states of a person. Here we use the language of mental states metaphorically, and, perhaps, in so doing try to change the meaning of the term. That is, to say that computers have mental states is to use "mental" in an extended sense. This strategy seems doomed to failure. It seems to blur rather than clarify what moral agency is.

Cognitive science is devoted to using the computational model to bring new understanding and new forms of knowledge. Cognitive scientists and computational philosophers seem to operate on the presumption that use of the computational model will lead to a revolutionary change in many fundamental concepts and theories.[13] To be sure, this promise has been fulfilled in several domains. However, when it comes to the debate over the moral agency of computers, the issue is not whether the computational model is transforming moral concepts and theories but whether a new kind of moral being has been created. In other words, it would seem that those who argue for the moral agency of computers are arguing that computers do not just represent moral thought and behavior, they *are* a form of it. After all, the claim is that computers do not just represent moral agency but *are* moral agents.

While this move from computational model to instantiation is not justified, the temptation to think of computers as more than models or simulations is somewhat understandable in that computers do not just represent, they also behave. Computer systems are not just symbolic systems; they have efficacy; they produce effects in the world, powerful effects on moral patients. Because of the efficacy of computers and computer systems, those who argue for the moral agency of computers are quite right in drawing attention to the moral character of computer systems. However, they seem to overstate the case in claiming that computer systems are moral agents. As will be discussed later, the efficacy of computer systems is

[13] For example, T. Bynum and J.H. Moor. *The Digital Phoenix: How Computers Are Changing Philosophy*. Basil Blackwell Publishers, Oxford, 1998. Bynum and Moor (1998) is devoted to describing how this has happened in philosophy.

always connected to the efficacy of computer system designers and users.

All of the attention given to mental states and non-determinism draws attention away from the importance of the intending to act and, more generally, away from intentionality. While computer systems do not have intendings to act, they do have intentionality and this is the key to understanding the moral character of computer systems.

The intentionality of computer behavior

As illustrated in discussion of the Cartesian doctrine, traditionally in moral philosophy, nature and machines have been lumped together as entities that behave mechanistically. Indeed, both nature and machines have been dismissed from the domain of morality because they have both been considered mechanistic. Unfortunately, this has pushed artifacts out of the sights of moral philosophy. As mechanistic entities, artifacts have been thought to be morally neutral and irrelevant to morality.

Because artifacts and natural entities have been lumped together as mechanistic, the morally important differences between them have been missed. Artifacts are human-made; they are products of action and agency. Most artifacts behave mechanistically once made, even though their existence and their design is not mechanistic. Artifact behavior, including computer behavior, is created, and used, by human beings as a result of their intentionality.

Computer systems and other artifacts have intentionality, the intentionality put into them by the intentional acts of their designers. The intentionality of artifacts is related to their functionality. Computer systems (like other artifacts) are poised to behave in certain ways in response to input. Johnson and Powers provide a fuller account of the intentionality of artifacts in which the intentionality of artifacts is connected to their functionality, and functionality is understood on the model of a mathematical function.[14] What artifacts do is receive input and transform the input into output. When, for example, using a search engine, I press certain keys entering particular words in the appropriate box and then press a button, the search engine goes through a set of processes and delivers particular output to my computer screen. The output (the resulting behavior) is a function of how the system has been designed and the input I gave it. The system designer designed the system to receive input of a certain kind and transform that input into output of a particular kind, though the programmer did not have to specify every particular output for every possible input.

In this way computer systems have intentionality. They are poised to behave in certain ways, given certain input. The intentionality of computer systems and other artifacts is connected to two other forms of intentionality, the intentionality of the designer and the intentionality of the user. The act of designing a computer system always requires intentionality – the ability to represent, model, and act. When designers design artifacts, they poise them to behave in certain ways. Those artifacts *remain* poised to behave in those ways. They are designed to produce unique outputs when they receive inputs. They are directed at states of affairs in the world and will produce other states of affairs in the world when used. Of course, the intentionality of computer systems is inert or latent without the intentionality of users. Users provide input to the computer system and in so doing they use their intentionality to activate the intentionality of the system. Users use an object that is poised to behave in a certain way to achieve their intendings. To be sure, computer systems receive input from non-human entities and provide output to non-human entities, but the other machines and devices that send and receive input and output have been designed to do so and have been put in place to do so by human users for their purposes.[15]

That computer systems are human-made entities as opposed to natural entities is important. Natural objects have the kind of functionality that artifacts have in the sense that they receive input and because of their natural features and composition, they transform input in a particular way, producing output. I pick up a stick and manipulate it in certain ways, and the stick behaves in certain ways (output). By providing input to the stick, I can produce output, e.g., collision with a rock. However, while both natural objects and human-made objects have functionality, natural objects were not designed by humans. They do not have intentionality. Most importantly, natural entities could not be otherwise. Artifacts, including computer systems, have been intentionally designed and poised to behave in the way they do – by humans. Their functionality has been intentionally created. By creating artifacts of particular kinds, designers facilitate certain kinds of behavior. So, it is with computers, though admittedly the functionality of computers is quite broad because of their malleability.

[14] Johnson and Powers 2005 (unpublished manuscript).

[15] This can also be thought of in terms of efficacy and power. The capacity of the user to do something is expanded and extended through the efficacy of the computer system, and the computer system exists only because of the efficacy of the system designer.

The point of this analysis of the intentionality of computer systems is twofold. First, it emphasizes the dependence of computer system behavior on human behavior, and especially the intentionality of human behavior. While computer behavior is often independent in time and place from the designers and users of the computer system, computer systems are always human-made and their efficacy is always created and deployed by the intentionality of human beings. Second, and pointing in a entirely different direction, since computer systems have built-in intentionality, once deployed – once their behavior has been initiated – they can behave independently and without human intervention.

The intentionality of computer systems means that they are closer to moral agents than is generally recognized. This does not make them moral agents because they do not have mental states and intendings to act, but it means that they are far from neutral. Another way of putting this is to say that computers are closer to being moral agents than are natural objects. Because computer systems are intentionally created and used forms of intentionality and efficacy, they are moral entities. That is, how they are poised to behave, what they are directed at, the kind of efficacy they have, all make a moral difference. The moral character of the world and the ways in which humans act are affected by the availability of artifacts. Thus, computer systems are not moral agents, but they are a part of the moral world. They are part of the moral world not just because of their effects, but because of what they are and do.

Computer systems as moral entities

When computer systems behave, there is a triad of intentionality at work, the intentionality of the computer system designer, the intentionality of the system, and the intentionality of the user. Any one of the components of this triad can be the focal point for moral analysis; that is, we can examine the intentionality and behavior of the artifact designer, the intentionality and behavior of the computer system, and the intentionality and behavior of the human user. Note also that while human beings can act with or without artifacts, computer systems cannot act without human designers and users. Even when their proximate behavior is independent, computer systems act with humans in the sense that they have been designed by humans to behave in certain ways and humans have set them in particular places, at particular times, to perform particular tasks for users.

When we focus on human action with artifacts, the action is constituted by the combination of human behavior and artifactual behavior. The artifact is effectively a prosthetic. The human individual could not act as he or she does without the artifact. As well, the artifact could not *be* and be *as it is* without the artifact designer (or a team of others who have contributed to the design and production of the artifact). The artifact user has a complex of mental states and an intending to act that leads to deploying a device (providing input to a device). The device does not have mental states but has intentionality in being poised to behave in certain ways in response to input. The artifact came to have that intentionality through the intentional acts of the artifact designer who has mental states and intendings that lead to the creation of the artifact. All three parts of the triad – the human user, the artifact, and the human artifact designer/maker have intentionality and efficacy. The user has the efficacy of initiating the action, the artifact has the efficacy of whatever it does, and the artifact designer has created the efficacy of the artifact.

To draw out the implications of this account of the triad of intentionality and efficacy at work when humans act with (and by means of) artifacts, let us begin with a simple artifact. Landmines are simple in their intentionality in the sense that they are poised to either remain unchanged or to explode when they receive input. Suppose a landmine explodes in a field many years after it has been placed there during a military battle. Suppose further that the landmine is triggered by a child's step and the child is killed. The deadly effect on a moral patient is distant from the landmine designer's intentionality both in time and place, and is distant in time from the intentionality of the user who placed the landmine in the field. The landmine's intentionality – its being poised to behave in a certain way when it receives input of a certain kind – persists through time; its intentionality is narrow and indiscriminate in the sense that any pressure above a certain level and from any source produces the same output, explosion.

When the child playing in the field steps on the landmine, the landmine behaves automatically and independently. Does it behave autonomously? Does it behave from necessity? Could it be considered a (im)moral agent? While there are good reasons to say that the landmine behaves autonomously and from necessity, there are good reasons for resisting such a conclusion. Yes, once designed and put in place, the landmine behaves as it does without the assistance of any human being and once it receives the input of the child's weight, it behaves of necessity. Nevertheless, the landmine is not a natural object; its independence

and necessity have been contrived and deployed by human beings. It is what it is and how it is *not* simply because of the workings of natural forces (though these did play a role). When the landmine explodes killing the child, the landmine's behavior is the result of the triad of intentionality of designer, user, and artifact. Its designer had certain intentions in designing the landmine to behave as it does; soldiers placed the landmine where they did with certain intentions. Yes, neither the soldiers nor the designers intended to kill *that* child, but their intentionality explains the location of the landmine and why and how it exploded.

It is a mistake, then, to think of the behavior of the landmine as autonomous and of necessity; it is a mistake to think of it as unconnected to human behavior and intentionality. To do so is to think of the landmine as comparable to a natural object and as such morally neutral. Landmines are far from neutral.

As already indicated, the landmine is, in terms of its functionality and intentionality, a fairly simple artifact. Yet what has been said about the landmine applies to more complex and sophisticated artifacts such as computer systems. Consider a computer system that is deployed to search the Internet for vulnerable computers, and when it finds such computers, to inject a worm.[16] The program, we can suppose, sends back information about what it has done to the user. We can even suppose that the program has been designed to learn as it goes the most efficient way to do what it does. That is, it has been programmed to incorporate information about its attempts to get into each computer and figure out the most efficient strategy for this or that kind of machine. In this way, as the program continues, it learns, and does not have to try the same complex series of techniques on subsequent computers. The learning element adds to the case the possibility that, over time, the designer and user cannot know precisely how the program does what it does. Moreover, the fact that the program embeds worms in systems means that it is not just gathering or producing information; it is "doing" something. The program has efficacy. It changes the states of computers and in so doing, causes harm to moral patients.

Does the added complexity, the ability to learn, or the wider range of input and output change the relationship between the system's intentionality and efficacy, and the intentionality and efficacy of the system designer and user as described in the case of the landmine? The answer is "no." Once designed and put in place, the program behaves as it does without the assistance of the person who launched it and behaves of necessity. Even when it learns, it learns as it was programmed to learn. The program has intentionality and efficacy. It is poised to behave in certain ways; it is directed at states of affairs in the world (computer systems with certain characteristics connected to the Internet) and is directed at changing those states of the world in certain ways. While designer and user may not know exactly what the program does, the designer has used his or her efficacy and intentionality to create the program and the user has deployed the program. When the program does what is does, it does not act alone; it acts with the designer and user. It is part of an action but it is not alone an actor. The triad of designer, artifact and user acted as one.

The fact that the designer and user do not know precisely what the artifact does makes no difference here. It simply means that the designer – in creating the program – and the user – in using the program – are engaging in risky behavior. They are facilitating and initiating actions that they may not fully understand, actions with consequences that they cannot foresee. The designer and users of such systems should be careful about the intentionality and efficacy they put into the world.

This analysis points to the conclusion that computer systems cannot *by themselves* be moral agents, but they can be components of moral agency. Computer systems (and other artifacts) can be part of the moral agency of humans insofar as they provide efficacy to human moral agents and insofar as they can be the result of human moral agency. In this sense, computer systems can be *moral entities but not alone moral agents*. The intentionality and efficacy of computer systems make many human actions possible and make others easier and therefore more likely to be performed. The designers of such systems have designed this intentionality and efficacy into them; users, then, make use of the intentionality and efficacy through their intentionality and efficacy.

Conclusions

My argument is, then, that computer systems do not and cannot meet one of the key requirements of the traditional account of moral agency. Computer systems do not have mental states and even if states of computers could be construed as mental states, computer systems do not have intendings to act arising from their freedom. Thus, computer systems are not and can never be (autonomous, independent)

[16] Technically this might simply be a program. The combination of program together with computers and the Internet (without which the program could not function) make it a system.

moral agents. On the other hand, I have argued that computer systems have intentionality, and because of this, they should not be dismissed from the realm of morality in the same way that natural objects are dismissed. Natural objects behave from necessity. Computer systems and other artifacts behave from necessity once they are created and deployed, but they are intentionally created and deployed. Our failure to recognize the intentionality of computer systems and their connection to human action tends to hide their moral character. Computer systems are components in moral action; many moral actions would be unimaginable and impossible without computer systems. When humans act with artifacts, their actions are constituted by their own intentionality and efficacy as well as the intentionality and efficacy of the artifact which in turn has been constituted by the intentionality and efficacy of the artifact designer. All three – designers, artifacts, and users – should be the focus of moral evaluation.

Since I argue against the moral agency of computer systems, why, one might wonder, do I bother to navigate through this very complex territory? To my mind, those who argue for the moral agency of computer systems accurately recognize the powerful role that computer systems play, and will increasingly play, in the moral character of the human world; they recognize that computer system behavior has moral character as well as moral consequences. Yet, while I agree with this, I believe that attributing independent moral agency to computers is dangerous because it disconnects computer behavior from human behavior, the human behavior that creates and deploys the computer systems. This disconnection tends to reinforce the presumption of technological determinism, that is, it reinforces the idea that technology has a natural or logical order of development of its own and is not in the control of humans. This presumption blinds us to the forces that shape the direction of technological development and discourages intervention. When attention is focused on computer systems as human-made, the design of computer systems is more likely to come into the sights of moral scrutiny, and, most importantly, better designs are more likely to be created, designs that constitute a better world.

References

B.R. Allenby. Engineering Ethics for an Anthropogenic Planet Emerging. In *Technologies and Ethical Issues in Engineering*, pp. 7–28. National Academies Press, Washington, D.C., 2004.

Aristotle. *Nicomachean Ethics*. Translation from Terence Irwin, Indianapolis, Hackett, 1985.

W.E. Bijker. Sociohistorical Technology Studies. In S. Jasanoff, G.E. Markle, J.C. Petersen and T. Pinch, editors, *Handbook of Science and Technology Studies*, pp. 229–256. Sage, London, 1994.

T.W. Bynum and J.H. Moor, editors, *The Digital Phoenix How Computers are Changing Philosophy*. Blackwell Publishers, Oxford, 1998.

J.H. Fetzer. *Computers and Cognition: Why Minds are not Machines*. Kluwer Academic Press, 2001.

L. Floridi and J. Sanders. On the Morality of Artificial Agents. *Minds and Machines*, 143: 349–379, 2004.

M. Heidegger. *The Question Concerning Technology and Other Essays*. Translated and with an Introduction by W. Lovitt. Harper & Row, New York, 1977.

T.P. Hughes. Technological Momentum. In L. Marx and M.R. Smith, editors, *Does Technology Drive History? The Dilemma of Technological Determinism*, pp. 12–12. The MIT Press, Cambridge, 1994.

D.G. Johnson and T.M. Powers. Computers as Surrogate Agents. In J. van den Hoven and J. Weckert, editors, *Moral Philosophy and Information Technology*. Cambridge University Press, 2006.

J. Law. Technology and Heterogeneous Engineering: The Case of Portuguese Expansion. In W.E. Bijker, T.P. Hughes and T. Pinch, editors, *The Social Construction of Technological Systems*, . MIT Press, Cambridge, 1987.

J. Pitt. *Thinking About Technology: Foundations of the Philosophy of Technology*. Originally published by Seven Bridges Press, New York, 2000.

B.C. Stahl. Information, Ethics, and Computers: The Problem of Autonomous Moral Agents. *Minds and Machines*, 14: 67–83, 2004.