Topoi (2013) 32:227–236
DOI 10.1007/s11245-012-9149-4

On the Moral Agency of Computers

Thomas M. Powers

Published online: 13 January 2013
© Springer Science+Business Media Dordrecht 2013

Abstract Can computer systems ever be considered *moral agents*? This paper considers two factors that are explored in the recent philosophical literature. First, there are the important domains in which computers are allowed to act, made possible by their greater functional capacities. Second, there is the claim that these functional capacities appear to embody relevant human abilities, such as autonomy and responsibility. I argue that neither the first (*Doman-Function*) factor nor the second (*Simulacrum*) factor gets at the central issue in the case for computer moral agency: whether they can have the kinds of intentional states that cause their decisions and actions. I give an account that builds on traditional action theory and allows us to conceive of computers as genuine moral agents in virtue of their own causally efficacious intentional states. These states can cause harm or benefit to moral patients, but do not depend on computer consciousness or intelligence.

Keywords Moral agency · Computer ethics · Machine ethics

Computer systems have taken on an amazing array of functions that were once exclusively the domain of human agency. In military affairs, finance, communications, transportation, law enforcement, medicine, and many other enterprises that affect human and environmental well being, computers do much of the work—that is, they are involved in the plans, decisions and actions in these areas. In many of these cases, computers are acting under the direct control of humans, acting semi-autonomously. But increasingly computers are programmed so that they carry out functions without the direct control or intervention of humans. They are effectively "taken off the leash" once held by human controllers, and are assigned important functions because of developments in capacities such as navigation, perception, diagnosis, and face-, object-, and speech-recognition (Moravec 2008). Some of these computers also have the ability to learn and interact with humans, other computers, and the world around them and thus to expand on the abilities given to them by their original programs. Other capacities that are on the technology horizon, such as proprioception, will only make computers' evolution towards agency more complete.[1] The trend is

Some of the views in this paper were presented as "Intentionality and Moral Agency in Computers" at the European Computing and Philosophy Conference in Pavia Italy, June 2004. I owe thanks to the commentators and participants. In addition, some of the views owe to joint work that I did with Deborah Johnson, including Johnson and Powers (2005a, b, 2008). The position I stake out here builds on these earlier views but departs significantly from them. I am indebted to Johnson and the many audiences that have responded to versions of this paper. All shortcomings of the present paper are my own.

T. M. Powers (✉)
Department of Philosophy, University of Delaware, Newark, DE, USA
e-mail: tpowers@udel.edu

T. M. Powers
Center for Science, Ethics, and Public Policy, University of Delaware, Newark, DE, USA

[1] Engineers at the EU-funded EPICS (Engineering Propioception In Computing Systems) state that "[p]roprioceptive computing systems collect and maintain information about their state and progress, which enables self-awareness by reasoning about their behavior, and self-expression by effectively and autonomously adapting their behavior to changing conditions" (www.epics-project.eu). Many skeptical philosophers will resist ascribing these abilities to computer systems, but it must be recognized that many members of the engineering community do not share this skepticism. The EU has funded five concurrent projects in "Self-Awareness in Autonomic Systems".

towards increasing the autonomy of computer systems, and with developments in robotics and nanotechnology, towards greater mobility and ubiquity. To engineers this autonomy is often seen in technical but not metaphysical terms—as indicating a relative independence from frequent human control. But to some philosophers computers are becoming a special kind of autonomous agent—they are becoming moral agents.

This account would have seemed fantastical 20 or 30 years ago, but some version of it is has now become accepted by many of those who do research in the field of computer ethics. In the recent literature, many authors have recognized the emerging case for artificial or computer moral agency as supported by the factors that I described above: (1) the increase in numbers and importance of domains in which computers act, made possible by greater technical (functional) capacities; and (2) the way in which these functional capacities, under proper philosophical interpretation, begin to embody the morally-relevant abilities that are found in human moral agents—autonomy, intentionality, responsibility, sensitivity to values, and the like. Let us call the first factor the *Domain-Function* view and the second the *Simulacrum* view—the latter indicating that most authors still believe that computers will remain inferior "copies" of the adult humans who are (indubitably?) full moral agents. The reason for separating out these two views will become apparent later; for now, I merely want to draw attention to the fact that Domain-Function considerations are primarily social and technical, whereas Simulacrum considerations are more narrowly philosophical.

The Domain-Function and Simulacrum views are often conjoined, and are almost always considered in some form in the accounts of computer moral agency in the literature. For instance, Johnson (2006) and Johnson and Miller (2008) are impressed by computers' increasing domain and function importance, but reject the "untethering" of computer systems from designers and the social conditions of their making and use, and thus reject computer moral agency. Sullins (2006), Moor (2006) and Wallach and Allen (2009) take a more optimistic view and, convinced by domain-function and simulacrum aspects of computers, are ready to accept (with various kinds of qualification) the moral agency of some computers—now or in the future. And in a series of influential papers, Luciano Floridi (2008) and his co-authors (Floridi and Sanders 2001, 2004; Floridi and Savulescu 2006) have argued for a sophisticated ontological framework of informational entities that can be understood—by means of a method of abstraction—as moral agents. They recognize (in fact, emphasize) what I have called domain-function. Yet Floridi and his co-authors insist that these entities *do not need* intentionality and responsibility in order to be moral agents. Their resulting view endorses the notion of "mind-less" moral agents.

I am impressed by this variety of views on the moral agency of computers, but at the same time a bit nostalgic for an account that would streamline the issue and return it to the traditional inquiry into the philosophy of action, where I think it belongs. For at the center of the controversy over computer moral agency is the question of whether the actions of a computer could be *of the sort that qualify* as moral actions as a result of the computer's decisions. In short, I worry that the treatment of this issue has tended to gloss over what must happen internal to the computer for it to be a moral agent when it acts in certain ways. I'd like to return the focus of the controversy to that issue—thorny as it may be.

Nonetheless, I am optimistic. In this paper I will outline an account of the moral agency of computers from that traditional standpoint: that computers will (someday soon?) act intentionally from having made reasoned moral decisions. Some of what I will describe in terms of computer system functional capacities can be seen as technological extensions of the already sophisticated computer systems of the sort that I mentioned above. Some of the account will be necessarily abstract, and its application will only become more concrete with future technological developments. Since I am not a futurist, I will not speculate on when all of these advances will occur, but I also do not think that an account of the moral agency of computers must await them. Basic theories from Aristotle to contemporary action theory are adequate to provide an outline of that account now.

1 Agency

There are two central, related concepts in human moral agency that I want to extend to the consideration of the moral agency of computers: (1) the ability to act on reasons, and (2) the having of intentional states. I will argue that a certain composite understanding of moral agency—what I call the "internalist" interpretation—will allow us to conceive of future computer systems as genuine moral agents because they will act on *moral reasons,* which reasons will be *their* reasons, and that by virtue of their intentional states they will be able to recognize and value real beings as separate from them. They will acknowledge that these separate beings have the status of moral patients, and thus they will be able to undertake actions to protect and maybe even to respect them. If I am correct in this speculation, the actions of a computer will thus be open to causal (reason) explanations in a way that is consistent with Davidson's theory of intentionality and is broadly compatible with much of contemporary action theory. We will be able to attribute a moral character to the actions of computers, and thus they will become moral agents. We will attribute this character to their actions not because

these actions will be a simulacrum of human moral agency, but because they will be moral actions proper.

Though this view is likely to appeal to cyber, AI, and sci-fi enthusiasts of many stripes—and surely the conclusion (if not the argument) has already occurred to many of them—such an extension of moral agency is not to be taken lightly and will be resisted by many philosophers who are proponents of a traditional philosophy of action. Granted, we should not be swayed by the "ordinary language" or pre-philosophical ascriptions of agency that are already common in computer science and engineering circles.[2] The internalist account of computer moral agency presented here is meant to be thoroughly philosophical. Moreover, I am not advocating a pragmatic or instrumental ascription of agency, as in Dennett's (1996) view, by suggesting merely that we adopt an "intentional stance" in order to interpret computers' behaviors. To the contrary, the explanations of their behavior on my account will owe to states that will be *their* reasons for acting. This is the gist of the internalist account of moral agency; an agent is someone or something that acts for reasons that may be influenced by external factors, but are nonetheless the agent's reasons.

A further caveat about this account of moral agency is in order. I will assume throughout that whatever it is that has allowed humans (through evolution) to become moral agents is entirely material, i.e., subject to the laws of physics. Philosophers' present understandings of this "something" may still be a kind of Lockean "we-know-not-what" support of agency and reason, though increasingly the neuroscientists believe they are drilling down to the details (Greene 2009). But whatever the finer details are of our brains and neurophysiology—the stuff that gives rise to reasons, intentions, and deliberations—there is no ghost; we are all machine. And if we are all machine, there is no necessary condition—no undesignable soul—that a computer system must lack, no ability that it cannot have in principle, without which it could not be a moral agent. Having set out this materialist plea for the possibility of computer moral agency, let us see what kind of positive case we are able to make for it.

We begin with some basic theories of agency. In the philosophical literature up through the 20th century there are four primary accounts: Causal, Aristotelian, Kantian, and Davidsonian views.[3] The overlapping components of these four accounts serve as cumulative additions to our contemporary understanding of moral agency.

Causal agency is of a sort that occurs frequently in cases at law. A single agent might be considered the sufficient cause or one of several singly necessary and jointly sufficient causes of an outcome such as change of state or even an injury. No intention, blame or responsibility is immediately attributed to causal agents, and indeed even though non-human sorts of causal agents such as earthquakes are "blamed" for deaths, this is not in a moral sense of blame. For a causal agent to be blamed in a moral sense, typically the violation of some norm is required.

Aristotelian agency produces voluntary acts that have deliberation beforehand; this agency requires cognition of the sort that is sensitive to facts (particulars), which for Aristotle is the role of *phronesis*. In attributing a kind of agency to deliberative human beings, Aristotle is extending the causal account and thus setting human rationality apart. As he says in Book III of the Nicomachean Ethics (2009) "Now every class of men deliberates about the things that can be done by their own efforts. For nature, necessity, and chance are thought to be causes, and also reason and everything that depends on man."

Kantian agency replaces deliberation with practical reason and introduces the notion of the will as a kind of causality that can effect action. Though affective elements in humans (our inclinations) and external pressures are always present and can be considered as kinds of "heteronomous" causality too, reason can be practical and hence can "serve as the determining ground of the will"—i.e., overrule the inclinations and other heteronomous influences. For Kant, it is only when practical reason achieves this dominion that we gain our autonomy. Hence Kant's view explicitly connects rational agency to a kind of metaphysical status, and in doing so also posits the human dignity of autonomous agents. This view also strengthens the internalist position of Aristotle; human individuals alone determine what practical reason demands, and the freedom-making activity of moral agency—Kant supposed—is entirely within one's control.

Davidsonian agency introduces the contemporary folk psychology and the everyday practice of "giving reasons" as an exercise in explaining one's own actions. Consider an action such as Smith embracing Jones. On the traditional causal account put forward by Davidson ([1963], 2001), several mental states of the agent Smith might constitute a primary reason and a causal explanation for the embrace. A primary reason consists of a "pro attitude of actions of a certain kind," along with other supporting beliefs that the action will in fact count as the relevant kind. For example, Smith might have a pro attitude about appearing warm and inviting, and he may believe that an embrace of Jones would let him appear as such. On Davidson's considered

[2] Not all of the technical engineering community ignores the philosophical issues in (what they call) agency. See, e.g., Davidsson and Johansson (2005).

[3] Johnson and Powers (2005a) identified a fifth kind of agency that issues from a triple of a technology, its designer, and the user; this composite agency gives rise to what we called Technological Moral Action. Johnson and Powers (2008) also identified a form of agency that applied more specifically to computer programs such as tax-preparation and even search-engine software: Surrogate Agency. In neither of these papers did we contend that computers are genuine moral agents.

view, an agent must have internal (mental) states that are intentional in the sense that they are about or are directed at other states of affairs. Examples might be a belief 'that it is now raining', or a fear 'of spiders'. Additionally, at least one of the intentional states must be a specific act of *intending*, i.e., what Bratman (1992) later includes under the concept of planning. Jointly, these intentional states constitute a reason for acting, and the agent can recite them as a *reason explanation* if so pressed. This account squares with our everyday experience of being asked such questions as 'Why did you shoot the robber?' On Davidson's view, responses that reveal intentional states, including 'I believed the robber was downstairs' and 'I feared being mugged by him', can give both the primary reason and the explanation of an action, i.e., your shooting the robber.

Though Davidson does not explicitly include further conditions for *moral* agency, they can easily be added in a way that is consistent with the Kantian, Aristotelian, and Causal views. Consider candidates for being a moral agent (A) and a moral patient (P). Let P be a moral patient if harm to P is morally relevant—on some theory of morality. Then A is a moral agent just in case A acts *from her own reasons* to cause harm or benefit to moral patient P. Thus, the internalist view of moral agency that we inherit from Davidson can incorporate the most important aspects of the prior views. Agents are causes of effects; they act after deliberating or through practical reason in a way that can be independent of inclinations; their internal states are about things outside of them and explain why they undertook certain actions with respect to those other things; and when the effects of their actions harm or benefit moral patients, they are moral agents.

My account of computer agency holds out hope that some computers, in principle, can act or have agency because their reasons are made of up their intentional states, and these states are causally connected to their behaviors. Computer agency becomes moral agency when these reason-caused behaviors have morally relevant effects on patients.

It will be helpful to keep in mind one obvious objection, though it can only be answered in due course. The objection comes in the claim that intentional states (and hence reasons) are nowhere to be found in the hardware and software of a computer, nor in any of their momentary configurations. As we shall see, this objection commits a basic ontological error in looking for reasons in the realm of physical (and electrical) states, as though they were the observables of a scientific experiment. This is a mistake in considering both computer and human agency, and all talk of reasons (or of intentional states) is part of descriptive folk psychology. The same "obvious objection" is easily turned against human agency, but to no avail. No specific configuration or state of neurons, synapses, or brain physiology could ever be *identified* with one's reasons for acting. Yet we continue to believe that human agents typically have reasons for acting. A reason is an abstraction outside of the material realm. But if we are materialists, we believe that the having of reasons (and Reason) depends on and emerges out of the physical/chemical/electrical states of the brain. If we are non-reductive materialists, we deny that reasons must be somehow re-assigned to brain states in order to be understood and to be causal components in action. There is no need to "explain away" the elements of folk psychology because they do not compete with our neuropsychology. Now, worries may arise in philosophy of mind and language over the reduction of speech acts to mental states and of mental states to brain states. But these issues will remain intractable until we consider that we are thinking about objects on different levels of explanation. The kind of explanation called for by morality (as a phenomena) is not the same as the kind required, for instance, by color blindness.

As should now be apparent, the plausibility of this view will depend to a large extent on whether we can make sense of a computer having reasons of it own, which in turn (on my view) depends on it having certain kinds of intentional states. Thus we turn to the consideration of intentional states.

2 Intentionality

The view that I am advancing is that computers, in principle, can genuinely act or have agency because they can have intentional states that are causally connected to their behaviors, and that these behaviors can have morally relevant effects on patients. But the case is even harder to make, since not any old ascription of intentionality will do. We need what I call internal intentionality in computers. In support of my view, I offer several examples to distinguish internal and external intentionality, and give reasons for thinking that computers can have both.

To return to the tradition: attempts to capture intentionality can be traced from theories of mind to theories of action, and through this bridge, to moral philosophy. As indicated above, the theory of action I am employing is a composite one suggested by Kant and Aristotle, and elaborated by Davidson (2001) and Searle (2001), among others. On this theory, in order for an action to be both open to "reason explanations" and subject to moral evaluation, there must be intentional mental states connected to an agent's action in some fairly specific ways. In fact, any modern moral theory in the rationalist tradition will make essential reference to intentional states. For Kant, the "right" way to move from intentional states to action is through the deliberative process of the Categorical

Imperative. This process requires an evaluation of maxims, which in modern terminology consist of complexes of beliefs, desires, goals, and plans. The relation of intentional states to actions, for Kant, is normative and not factual.

How then are we to understand the action/intention connection? Consider this example, drawn from similar examples given by Davidson. Suppose Smith has a particular desire to kill Jones. On the same explanatory level as the desire is the corresponding description of it as the intentional state consisting of the psychological attitude 'the desire for killing X' and the content picked out by the term 'Jones'. Surely, Smith will have other intentional states as well, such as beliefs about the location of the murder weapon and the strength of the requisite blow to Jones that will kill him. Smith will also have (in the moment) the intention-in-action of swinging the weapon down on Jones' head. Absence of some of these intentional states in Smith—in the event, let us say, that he had no desire to kill Jones, but rather tripped and accidentally struck Jones with a blunt object—would alter the "reason explanation" of the event and (probably) negate an ascription of moral agency. (I hedge here because Smith may have been grossly negligent in the way in which he was carrying the blunt object). Now a similar scenario might unfold if Jones' death had been caused by a computer. If the intentional states of the computer caused some lethal behavior to be executed by it, and this behavior killed Jones (not because of a bug but in virtue of the "normal" operation of the software) we would be tempted to attribute moral agency to the computer as a whole. In this story, we would have to separate out the several representations in the computer, e.g., 'at coordinates <p,q,r> stands Jones'; 'Jones is to be killed'; 'execute lethal force against Jones'. Similarly, we would not attribute agency if the program malfunctioned because an electrical surge switched a series of gates in a way that was not controlled by the software. Notice here that the "intelligence" of the computer is not really at issue; what we need to know at this point is whether the computer's intentional states caused harm to a moral patient.

Now, granting for the moment that the intentional states of the computer did cause its lethal behavior, what about the disclaimer these states were merely a function of its program? Don't we "defend" the computer here by pointing out that it operates on representations of rules and commands (in the form of algorithms) that were created externally to the computer, and later imposed on it? Don't we believe that the computer "couldn't have done otherwise"? To answer these questions, we need to look closer at the distinction between internal and external intentional states.

Following Dretske's well-known example of the thermostat (1980), I claim that all designed artifacts have a kind of generic "aboutness" or intentionality. Other philosophers have offered similar observations, including Searle (1983) in his account of observer-relative intentionality. But our account of agency will require a kind of intentionality that is "of the computer." Haugeland's (1990) distinction between original and derivative intentionality comes closer to properly capturing the aspects of intentionality that are most relevant to the kind of action under consideration. We need to explore the kinds of intentionality in various guises: in human minds, in expressions such as sentences and speech acts, and (crucially) in representational states that are found in (and can be produced by) computers.

As befits my internalist account of moral action, the distinction I favor is that between internal and external intentional states. Internal intentional states have been claimed to be those that *necessarily* remain mental while external states are expressed in forms outside of the body, such as speech acts, written sentences, maps, and other designed artifacts. These artifacts include computer systems, but I will not assume (contrary to the tradition) that computers are only capable of external intentional states.

Much of the recent literature on computer moral agency considers intentionality, but does not distinguish between internal and external kinds. Perhaps this is so because computers manifestly *do* express external intentional states, and no doubt early digital computers were capable of nothing more. A display might represent, through symbols, a warning such as 'Kernel panic!' A software model might represent the ocean tides, or even something more complex like measurements of ocean acidification. Unless the skeptic doubts all forms of intentionality, he or she should readily admit that the symbols produced by a human-created program are intentional in this external sense. Computers work (in part) because they are machines that run on the back of proof theory. Their operations are truth preserving, just as is a valid deduction written on a chalkboard. But how is it that computers are meaning preserving? One answer is that what is generated by a computer, in the symbol system of a variety of languages, has meaning for humans precisely because humans have put a kind of meaning-preserving functionality into the programs. All of these symbols are external intentional states originally generated by and subject to the interpretations of human computer builders and programmers.

This is not to say that a computer is a mere repeater of strings of symbols that are programmed into it—that its output is merely a version of the input of human programmers. A computer is indeed more than a typewriter—a technology that is limited to reproduce input as output of a different form. While a computer can produce novel strings of symbols according to syntactic rules or other transformation rules, it has been argued—most forcefully by Searle

(1980) in his famous Chinese Room paper—that computers cannot produce a semantics to interpret those strings of symbols, and thus will never be capable of having internal intentional states of their own. All "meaning" in a computer is introduced from without. Of course, there were (and still are) many critiques of the Chinese Room argument, some of which envisioned ways for computers to have "meanings of their own." One of Searle's challenges to his critics allows that, even assuming a computer could produce its own semantics, we have no reason to expect the output of such a computer to be understandable to us. So why is it that computers produce intentional states (outputs in English sentences on screens) that we understand so very well? Is this just a coincidence, or are the computers providing a bridge from their semantics to ours, just for our benefit? Searle believes that such questions bring doubt on the prospect of Strong AI, and suggest that if a computer were to become intelligent, we might not understand *it*.

But let us take a step back here and consider just what is up for debate, for our account does not require all aspects of a thinking or intelligent computer. External intentionality is ubiquitous in written symbols, speech acts, and the forms and functions of all designed artifacts.[4] To design an artifact, the designer must start with a specification, based on a model of the world and the users' capabilities. The characteristics of the artifact must be "about" these things that are external to the artifact; design makes essential reference to things in the world, so to speak. These representations of users and environments, while not syntactically structured, reside in some form in all artifacts. Engineers know that the things they design do not somehow magically take shape, form, and function; they must be planned and constructed. Now, the world gives the designer guidance, as design is grounded in facts such as the shape and size of the human hand—which generates the shape of the shovel handle in a good design. Likewise, design is grounded in norms; the sloped pitch of a wheelchair-accessible sidewalk ramp ought to take account of the way in which both mobility-able and mobility-impaired users typically get around.

The external intentionality of designed artifacts is revealed precisely when the design is bad—when the shovel doesn't fit the hand, or when the sidewalk prevents mobility as opposed to facilitating it. As Haugeland (1990) rightly notes, "the fallibility of intentionality reveals that it is not merely a factual but also a normative relation." So I take it that for the purposes of my account, the existence of external intentionality in designed artifacts is the more straightforward claim to establish. But how could sophisticated designed artifacts, such as some computer systems, have something more—what I've called internal intentionality? And how could we recognize *internal* intentionality when it derives from a source other than the human mind?

One way of testing for internal intentionality relies on the notion of meaning and the first-person access to mental states. I understand my own thoughts, beliefs, and desires, because I have—for the most part—immediate access to them. While computer systems can generate external intentional symbols (a sentence in English on the display, for instance), it may appear to some that these states can be meaningful only to humans. This is supposedly the strong claim in Searle's argument, but it misses the point of the internal/external distinction. We are not concerned here with sentences that the computer generates for our (the user's) benefit, but with internal states that it has which emerge from (but are not identical to) the states of its hardware and software. The position I am advocating about what is internal to the computer would be the counterpart to a non-reductive materialism about human mental states.

So now let us turn the issue of first-person access on its head. What evidence could we possibly have to justify the claim that *these* internal states of a computer are not meaningful to it—that they are *not* the states that caused it to act in a certain way? I can think of none. Moreover, the only evidence that one human can have of what caused another human to act must be expressed as external intentional states—verbal or written reports, or other signs or symbols, generated by the putative agent. So the idea that, in order to establish computer moral agency, we must somehow investigate the internal states of a computer to find out if they are intentional (and in fact that they caused the action) suggests a standard that cannot be upheld when we go to consider human moral agency. The only causal explanations that any human will ever have that allow direct reference to internal intentional states are explanations of one's own behavior!

A different thought experiment might be illuminating in the comparison of human and computer agency. Suppose that we want to understand a human test subject's aberrant behavior. We might place the subject's head in an fMRI machine and read off his brain states.[5] Supposing that we were well-trained in neuroscience, the resulting graphical representations might mean something to us, but not to the subject. Suppose we tell the subject that what we've learned explains the aberrant behavior. Our account would miss the point by virtue of the explanatory mistake. The

[4] This last claim is the central thesis of Johnson and Powers (2005b), where it was argued that we ought to view all technological artifacts as carrying the (external) intentional states of their designers. The position I'm articulating here goes well beyond that claim: that certain artifacts (computers) can have, in addition, their own intentional states.

[5] Actually, all the fMRI does is to approximate brain activity through a measurement of the blood-oxygen levels of cells.

test subject's brain states are *not his* reasons for the aberrant behavior. Those reasons—to him and to others who try to understand him—traffic in the explanatory language of beliefs, desires, and plans, and not in the science of neurophysiology and physics.

3 AI and Moral Agency

On the optimistic side of the debate about what computers can and cannot do is the AI community. The literature in AI seems particularly receptive to the internalist view of computer agency. For the last several decades AI enthusiasts (Brooks 2002; Kurzweil 2000; Danielson 1992) as well as moral philosophers interested in information technology (Allen et al. 2000; Floridi and Sanders 2001, 2004) have discussed the prospects of "artificial" or non-human agency. Artificial agency can be considered alongside another late 20th Century theory: the computational theory of mind (CTM). This theory dominated the latter part of 20th Century philosophy of mind. How does CTM support the internalist account we are considering?

Recall that intentionality is "aboutness": the property of a process, state or entity such that it is directed at or represents another entity or state of affairs. Many adherents of CTM believe that human cognition can be both computational and essentially intentional or representational. The ability of formal languages to represent, and of computers to manipulate representations, does suggest a compelling alignment of intentional cognition with electro-mechanical computation.

According to Fodor (2000), "intentional processes are syntactic operations defined on mental representations" and thus many (if not all) of our cognitive processes are essentially "structured mental representations that are much like sentences." This comparison of cognition (by means of intentional states) to computing (by means of representations) allows Fodor to see the state changes of a computer as tracking the state changes in the human mind. With slight variations, these views are widely shared by the believers in the CTM (McClintock 1995).

The CTM does have implications for studies of agency as well as cognition. According to Haugeland (1990), the CTM holds that intentional ability depends on the thing being a "semantically interpreted active syntactical system" and one such implementation is a digital computer. Something cannot have original intentionality, according to Haugeland, "unless it can think about itself, and its own thoughts, or about the thoughts of others, or about the difference between truth and error, or about norms and values." Since (on the view I have been defending) norms and values are *about* moral patients that are in the world, and because they are separate from the moral agent, these patients must be represented by the agent in order to become part of its reasons for acting. So in consideration of what I claim about agency, Haugeland's conditional above must also go the other way: the agent cannot think about norms and values unless it can have original intentionality.[6]

When we combine the CTM with our previous account of agency, we start to see a stronger case for the moral agency of computers. If human minds have intentional states, in virtue of which they act, and computers operate on representations that might one-day mimic the logical structure of human mental processes (including human intentional states), what is the *moral* difference between the agency of a human and the agency of a computer? Here we see why we must dispense with the Simulacrum view. Computers will be moral agents only if they have genuine internal intentional states.

Herbert Simon (1977) claims that we have had, for some time now, such artificially intelligent computers capable of producing their own forms of intentionality and thus capable of reasoned action in the robust sense of traditional action theory. Two examples of intelligent systems that exhibit their own intentionality, according to Simon (1996), are the robotic self-navigating vehicle and the chess-playing computer. The former features a program that "[has] the *intention* of proceeding along the road and remaining on it [and] creates internal symbols that denote [landscape] features, *interprets them*, and uses the symbols to guide its steering." Likewise, the chess robot "forms an internal (symbolic) representation of the chess position. The symbols in this internal representation denote the external physical pieces and their arrangement, and the program demonstrates quite clearly, by the moves it chooses, that it *intends* to beat its opponent."

Allen (1995) compares the alleged production of intentional mental states in animals to those in humans, and urges an inclusive view of what counts as intentionality. Allen et al. (2000) suggest that there might be several ways to build an artificial and autonomous moral agent. These inclusive views on agency suggest that the representational states of an artificially intelligent computer should not be prejudiced as merely "artificial" or derived forms of intentionality.

So in answer to the question above: I can find no moral difference in the agency of the human and a computer of the right (intentional) sort. Still, there are differences aplenty in their agency; one of them is consciousness. And here I am convinced by Floridi's conclusion of "mind-less morality," though for reasons that do not depend on his framework of informational entities. Much contemporary

[6] Now, whether original and internal intentionality diverge might be a question to be answered for each type of candidate agent. But I see no reason why they must diverge, and I don't think this matters for my view.

work on consciousness focuses on its contents, some of which I have been calling internal intentional states: beliefs, desires, plans, and states of intending. The "contents of consciousness" are these very states (along with other non-intentional states). The intentional states themselves have contents—they make reference to things outside of the mind—and thus consciousness might be said to be "doubly" content-ful. But this condition is not a necessary condition of having intentional states; proof of this is apparent in the fact that a book can contain intentional entities and yet not be conscious. It is, indeed, a remarkable property of the human mind that it can have contents in this way—consciously—but surely this cannot be a requirement of all content-having entities.

Wallach and Allen (2009) also forego any necessary attribution of consciousness to computer moral agents, but do so for pragmatic reasons. They recognize the technical and philosophical difficulties in programming consciousness into computers, and insist that "functional equivalence of behavior is all that can possibly matter for the practical issues of designing artificial moral agents." What I would urge here is that designing computers so that they have their own intentional states is necessary for the functional equivalence of their behaviors to human behaviors. But I too am more than willing to dispense with consciousness as a condition for moral agency.

Even though I have here discussed (briefly) only CTM, I do not intend to weigh in on the debate over computationalist, connectionist, or dynamicist views of cognitive science, nor to characterize the manner in which intelligence might be represented in computers. All of these views, at some level of explanation, are committed to there being representational states in the computer and in human minds (Symons 2001). I need not exploit what they agree upon—that an artificially intelligent computer must have representational states—in order to argue that a computer moral agent must have representational states. It is true that, on functionalist and computationalist views of cognitive science, computers could be or become intelligent beings insofar as their computer operations simulate or mimic rational thought in humans. But this goes farther than we need to go. Whether these "arguments from simulation" are unfounded, they are unnecessary to make the case for moral agency. The concept of moral agency is elastic enough to allow that different kinds of entities can act in a moral sense. Computers need not simulate humans or their intelligence in order to be moral agents.

4 The Moral Status of Computer Agents

Readers who are convinced so far may still balk at extending moral agency to computers for fear that we will owe them something—perhaps a certain kind of treatment or level of respect—if we acknowledge their agency. The question of what special moral concern, if any, is owed to computers as agents brings up again the useful distinction between moral agents and moral patients.

Recall that according to the definition of moral agency, it does not follow that everything that is an agent is also a patient. Likewise, not all patients are agents. This latter claim is most obvious in the consideration of small children, who lack the rational capability for full agency (though they at least have intentional states, we suppose). It would be very strange to say that small children are moral agents, and certainly we do not treat them as such, but they are clearly moral patients.

Whether something is a moral patient—whether harm to it matters morally—will depend on the theory that one holds. For utilitarians, only sentient beings will count. For strict Kantians, only "transcendentally free" human agents with a noumenal soul, and for contemporary deontologists only rational autonomous beings. For Aristotelians, only humans with a biologically determined purpose in life matter morally. For some environmental ethicists, entities such as species and ecosystems might count. It is perhaps too much to require a new moral theory just to accommodate artificial agents, but it is worth noting that on any of the theories just considered, computers are unlikely to be moral patients.

In considering the logical relations between classes of moral patients and agents, I think Floridi and Sanders (2004) were too quick to reject as unrealistic the possibility that the class of moral agents and patients might only intersect, thus leaving some moral agents that are not moral patients. This is the situation, I believe, with computers that are moral agents. Without knowing more about their abilities to feel pain, or possess dignity, or even have purposes that are themselves morally desirable, I am unconvinced that we must acknowledge their moral worth. Denying this moral status does not alter their moral agency, on the view that I've put forth, since nothing about being a moral patient would seem to include the ability to have internal intentional states that count as reasons for action.

One possibility is that the Simulacrum view can actually be of use here—on the issue of moral patients, instead of moral agents. I could imagine that an artificially intelligent computer would be of special interest and warrant a kind of moral respect insofar as it had *moral intelligence*. For instance, it seems that a computer would be morally intelligent if it could exhibit grief at the right moment, feel guilt appropriately, worry about others' happiness, and maybe even contemplate notions like dignity, respect, and trust. But the list of moral theories and accounts of patients mentioned above should give us pause; do we really know what it is about entities that makes them morally

estimable? There is great variation over theories concerning the necessary qualities of a patient, and even if we reach a consensus on what kinds of things are moral agents, and that consensus includes certain kinds of computers, I suspect the issue of patients will remain divisive. Besides, we have reasons to distrust the Simulacrum view as applied to moral patients. Similarity or likeness to "standard" moral patients in the past has been used as a pretext to discriminate against women and minorities. Other comparisons to the "standard" have convinced some people to disregard the moral status of many of the quite intelligent and social higher primates.

5 Conclusion

I have presented an account of the moral agency of computers that is partly speculative, but also conservative in the sense that it does not require a radical revision of our contemporary (traditional) understanding of agency. It does however alter our traditional conception of the class of agents. I suspect that, if I am right, I will not be shown to be right with some sudden advance in programming or hardware. Progress in the moral abilities of computers, as I (Powers 2011) and Wallach and Allen (2009) have argued, is incremental and should come from engineers responding to increased computer functionality and (alas) lethality. In this respect, the Domain-Function considerations seem correct. What computers can do, and how much they are allowed to do, seem to be factors that "run along side of" our judgments of moral agency.

There is little reason to think that computer engineers will directly design moral agency into computers as though it were a kind of software module. Rather, moral agency is likely to be reached by continuing on the path that we are now on. This path includes, as I mentioned at the start of this paper, amazing advances in sensing, mobility, and awareness of surroundings. It also includes the ability to recognize other agents and value patients—hence the ability of intentionality—and this is the crucial step on the way to moral agency.

Where the Domain-Function considerations can lead us astray is when they "get in front of" the moral agency of computers. What I have in mind are the kinds of economic and "efficiency" arguments that can be used to deploy computers in domains before they are able to do the assigned tasks with a level of safety and moral discernment that would be expected of the best human moral agents. These arguments also facilitate an over-estimation of functional capacities.

So in decisions about deploying advanced computer systems, the human decision makers may find themselves in a predicament like that of an employer hoping to hire an employee for a completely new position. Prior "domains" in which an applicant has worked are relevant, as are any amazing, demonstrated abilities of the applicant. But that an applicant has had many other important jobs and demonstrates amazing skills is no guarantee that they can do the new job. Likewise, the Domain-Function considerations that sway many computer ethicists do not go, directly, to the question of moral agency—not at least on the traditional view that I have here advocated. And when these considerations precede development of the moral agency of a deployed computer in a critical context, I fear we will learn a hard lesson.

A lingering question about my account will be how it can be that, though computers are programmed with representations that are the *external* intentional states of their human programmers, yet somehow they will produce *internal* intentional states as the constituents of their own reasons for acting. In sketching an answer to this puzzle, perhaps an analogy will help. Let us suppose that a parallel situation exists in typical cases of human reproduction. In some sense, parents contribute the DNA to their child, from which it gains its genetic inheritance and a blueprint for its development. This will be its "software and hardware." Much of that development is left undetermined, especially the particulars of the child's own upbringing. Nonetheless, the lessons of the child's upbringing could also be represented as instructions, given to the child from "external" sources. What the child experiences will play some role in its agency, and along with its DNA will inform the reasons it will have (at any moment) for acting.

Will the child ever have intentional states of its own? And will these states, along with other cognitive and physical abilities, allow it to act for moral reasons? There is overwhelming evidence that this happens. Future computers are likely to have, in many respects, the same abilities, and how they learn and what they experience will influence their behavior. They will never operate independently of their blueprint, nor I suspect will they be entirely determined by it.

References

Allen C (1995) Intentionality: natural and artificial. In: Meyer JA, Roitblat HL (eds) Comparative approaches to cognitive science. MIT Press, Cambridge, pp 93–110

Allen C, Varner G, Zinser J (2000) Prolegomena to any future artificial moral agent. J Exp Theor Artif Intell 12:251–261

Aristotle (2009) The Nicomachean ethics, (trans. WD Ross,). Oxford University Press, Oxford

Bratman ME (1992) Planning and the Stability of Intention. Minds Mach 2(1):1–6

Brooks RA (2002) Flesh and machines: how robots will change us. Pantheon Books, New York

Danielson Peter (1992) Artificial morality: virtuous robots for virtual games. Routledge, New York
Davidson D (1963) Actions, reasons, and causes. J Philos 60:685–700. Reprinted in Davidson (2001) Essays on actions in events. Oxford University Press, Oxford, pp 3–20
Davidsson P, Johansson S (2005) On the metaphysics of agents. In: Proceedings of the fourth international joint conference on autonomous *agents and multiagent systems*, pp 1299–1300
Dennett DC (1996) The intentional stance. MIT Press, Cambridge
Dretske FI (1980) The intentionality of cognitive states. In: French P et al (eds) Midwest studies in philosophy, vol 2. University of Minnesota Press, Minneapolis, pp 281–294
Floridi L (2008) Information ethics: its nature and scope. In: van den Hoven MJ, Weckert J (eds) Information technology and moral philosophy. Cambridge University Press, Cambridge
Floridi L, Sanders JW (2001) Artificial evil and the foundation of computer ethics. Ethics Inf Technol 3:55–66
Floridi L, Sanders JW (2004) On the morality of artificial agents. Minds Mach 14:349–379; also in Anderson and Anderson (2011)
Floridi L, Savulescu J (2006) Information ethics: agents, artefacts and new cultural perspectives. Ethics Inf Technol 8:155–156
Fodor J (2000) The mind doesn't work that way: the scope and limits of computational psychology. MIT Press, Cambridge
Greene JD (2009) The cognitive neuroscience of moral judgment. In: Gazzaniga M (ed) The cognitive neurosciences, 4th edn. MIT Press, Cambridge
Haugeland J (1990) The intentionality all-stars. Philos Perspect 4:383–427. Reprinted in Haugeland J ed (1998) Having thought. Harvard University Press, Cambridge, pp 127–170
Johnson DG (2006) Computer systems: moral entities but not moral agents. Ethics Inf Technol 8:195–204; also in Anderson and Anderson (2011)
Johnson DG, Miller K (2008) Un-making artificial moral agents. Ethics Inf Technol 10:123–133
Johnson DG, Powers TM (2005a) Computer systems and responsibility: a normative look at technological complexity. Ethics Inf Technol 7(2):99–107, Kluwer
Johnson DG, Powers TM (2005b) Ethics and technology: a program for future research. In: Mitcham C (ed) Encyclopedia of science, technology, and ethics. Macmillan Reference, Detroit
Johnson DG, Powers TM (2008) Computers as surrogate agents. In: van den Hoven MJ, Weckert J (eds) Information technology and moral philosophy. Cambridge University Press, Cambridge
Kurzweil Ray (2000) The age of spiritual machines. Viking Penguin, New York
McClintock Alexander (1995) The convergence of machine and human nature. Ashgate, Hampshire
Moor JH (2006) The nature, importance, and difficulty of machine ethics. IEEE Intell Syst; also in Anderson and Anderson (2011)
Moravec H (2008) Rise of the robots. Sci Am 18:12–19
Powers TM (2011) Incremental machine ethics. IEEE Robot Autom 18(1):51–58
Simon HA (1977) Models of discovery: and other topics in the methods of the sciences. Springer, New York
Simon HA (1996) Machines as mind. In: Millican P, Clark A (eds) Machines and Thought: The Legacy of Alan Turing, vol 1. Oxford University Press, Oxford
Searle J (1980) Minds, brains, and programs. Behav Brain Sci 3:417–424
Searle J (1983) Intentionality. Cambridge University Press, Cambridge
Searle J (2001) Rationality in action. MIT Press, Cambridge
Sullins JP (2006) When is a robot a moral agent?. Int Rev Inf Ethics 6:23–30; also in Anderson and Anderson (2011)
Symons J (2001) Explanation, representation, and the dynamical hypothesis. Minds Mach 11:521–541
Wallach W, Allen C (2009) Moral machines: teaching robots right from wrong. Oxford University Press, New York