Ethics and Information Technology (2009) 11:19–29
DOI 10.1007/s10676-008-9167-5

Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?

Kenneth Einar Himma
Department of Philosophy, Seattle Pacific University, 3307 Third Avenue West, Seattle, WA 98119, USA
E-mail: himma@spu.edu

Abstract. In this essay, I describe and explain the standard accounts of agency, natural agency, artificial agency, and moral agency, as well as articulate what are widely taken to be the criteria for moral agency, supporting the contention that this is the standard account with citations from such widely used and respected professional resources as the *Stanford Encyclopedia of Philosophy*, *Routledge Encyclopedia of Philosophy*, and the *Internet Encyclopedia of Philosophy*. I then flesh out the implications of some of these well-settled theories with respect to the prerequisites that an ICT must satisfy in order to count as a moral agent accountable for its behavior. I argue that each of the various elements of the necessary conditions for moral agency presupposes consciousness, i.e., the capacity for inner subjective experience like that of pain or, as Nagel puts it, the possession of an internal something-of-which-it is-is-to-be-like. I ultimately conclude that the issue of whether artificial moral agency is possible depends on the issue of whether it is possible for ICTs to be conscious.

Key words: accountability, agency, artificial agents, consciousness, ethics, moral agency, natural agents

Introduction

A spate of papers has recently appeared on the possibility of artificial agency and artificial *moral* agency, raising substantive questions of whether it is possible to produce artificial agents that are morally responsible for their acts. As ICTs become more sophisticated in their ability to solve problems, a host of issues arise concerning the moral responsibilities for the acts of ICTs sophisticated enough to raise the possibility that they are moral agents and hence morally accountable for their acts.

In this paper, I will work out the details of the standard accounts of the concepts of agency, natural agency, artificial agency, and moral agency, as well as articulate the criteria for moral agency. Although the claims I rely upon are so widely accepted in the philosophical literature that they are taken for granted in such widely used and respected professional resources as the *Stanford Encyclopedia of Philosophy*, *Routledge Encyclopedia of Philosophy*, and the *Internet Encyclopedia of Philosophy*, I will explain the rationale for these claims – and why they are widely regarded as uncontroversial. Although there are a number of papers challenging the standard account, I will not consider them here. My focus is on working out the implications of the standard account; an evaluation of the non-standard accounts could not adequately be done in the space available here.

I will begin with analyses of the more basic concepts, like that of agency, and work up to analyses of the more complex concepts, like that of moral agency, subsequently considering the meta-ethical issue of what properties something must have to be accountable for its behavior. I will then argue that each of the various elements of the necessary conditions for moral agency presupposes consciousness, i.e., the capacity for inner subjective experience like that of pain or, as Nagel puts it, the possession of an internal something-of-which-it-is-like-to-be and that the very concept of agency presupposes that agents are conscious. I ultimately conclude that the issue of whether artificial moral agency is possible depends on the issue of whether it is possible for ICTs to be conscious.

The concept of agency

The idea of *agency* is conceptually associated with the idea of being capable of doing something that counts as an act or action. As a conceptual matter, X is an agent if and only if X is capable of performing actions. Actions are *doings*, but not every doing is an

action; breathing is something we do, but it does not count as an action. Typing these words is an action, and it is in virtue of my ability to do this kind of thing that, as a conceptual matter, I am an *agent*.

It might not be possible for an agent to avoid doings that count as actions. Someone who can act who chooses to do nothing, according to the majority view, is doing something that counts as an action – though in this case it is an omission that counts as the relevant act. I have decided not to have a second cup of coffee this morning, and my ability to execute that decision in the form of an omission counts, in a somewhat technical sense, as an act, albeit one negative in character. Agents are not merely capable of performing acts; they inevitably perform them (in the relevant sense) – sometimes when they do nothing.

The difference between breathing and typing words is that the latter depends on my having a certain kind of mental state, while the former does not. Some theorists, like Davidson, regard the relevant mental state as a belief/desire pair; on this view, if I want X and believe *y* is a necessary means to achieving *x*, my belief and desire will cause my doing *y* – or will cause something that counts as an "intention" to do *y*, which will cause the doing of *y*. Others, including myself, regard the relevant mental state as a "volition" or a "willing."[1] For example, if I introspect my inner mental states after I have a made a decision to raise my right arm and then do so, I will notice that the movement is preceded by a somewhat mysterious mental state (perhaps itself a doing of some kind) that is traditionally characterized as a "willing" or a "volition." Either way, it is a necessary condition for some event *y* to count as an action that *y* be causally related to some other mental state than simply a desire or simply a belief.

Breathing is not an action precisely because my taking a breath at this moment doesn't depend directly on an intent, belief/desire pair of the right kind or volition – though it might depend indirectly on my not having a particular mental state, namely an intention to end my life. Waking up in the morning is something I do, but it is not an action, at least not most of the time, because it doesn't involve one of these conscious states – though getting out of bed does.

The relevant mental states might be free or they might not be free. Volitions, belief-desire pairs, and intentions might be mechanistically caused and hence determined – or they might be free in some libertarian or compatibilistic sense. Likewise, the relevant mental states might be related to something that counts as the kind of mental calculation (e.g., a deliberation) that we associate with rational beings – or they might not be.

Agency is therefore a more basic notion than the compound concepts of free agency, rational agency, and moral agency – although it may turn out that one must be rational or free to count as an agent. Ordinary intuitions about the relevant concepts differ. On the one hand, it might be thought that one need not be either rational or free to be an agent. While dogs are neither rational nor free (in the relevant sense), it makes sense to think of them, on this view, as being capable of performing actions because some of their doings seem to be related to the right kinds of mental states – states that are *intentional* in the sense that they are about something else but not necessarily in the sense of having an intent or intention (which seems to presuppose linguistic or quasi-linguistic abilities absent in dogs).[2] One the other, it might be thought that intentional states can be instantiated only by beings with linguistic capabilities. Many theorists, for example, believe that dogs cannot have beliefs, although they behave in ways that suggest they have correlative mental states, because belief is a matter of assenting to propositions and a dog cannot understand a proposition and hence cannot assent to it. I will not take a position on this issue, as nothing of importance turns on it.

Only beings capable of intentional states (i.e., mental states that are *about* something else, like a desire for X), then, are agents. People and dogs are both capable of performing acts because both are capable of intentional states; people are, while dogs might not be (if the first of the two lines of argument in the last paragraph is sound), *rational* agents because only people can deliberate on reasons, but both seem to be *agents*. In contrast, trees are not agents, at bottom, because trees are incapable of intentional states (or any other mental state, for that matter). Trees grow leaves, but growing leaves is not something that happens as the result of an action on the part of the tree.

Agency, as a conceptual matter, is simply the capacity to cause actions – and this requires the capacity to instantiate certain intentional mental

[1] I want to remain agnostic with respect to theories of mind. A mental state might be a private, inner state that is non-extended and non-material, as substance dualism and non-reductive physicalist theories assert, or it might be nothing more than a brain state, as reductive physicalism (e.g., identity theory) asserts. I make no assumptions here about the nature of a mental state generally.

[2] See Pierre Jacob, "Intentionality," *Stanford Encyclopedia of Philosophy* (Edward Zalta, ed.); available at http://plato.stanford.edu/entries/intentionality/.

states. Usually, it is thought that these mental states are the cause of the action, but there is some controversy in philosophy of mind over whether mental states are epiphenomenal or not (i.e., incapable of causing actions or anything else). While most philosophers of mind believe that mental states play a causal role in our behavior, not all do – and those that do not may disagree over whether human beings would count as agents if it is neurophysiological states doing all the causal work while the associated mental states are epiphenomenal. The most common view, and one I will assume here without defense, is that it is a necessary condition of agency that the relevant *mental* states are capable of causing performances (though, again, I make no claims here about exactly what that state is). Thus, the following constitutes a rough but accurate characterization of the standard view of agency: X is an agent if and only if X can instantiate intentional mental states capable of directly causing a performance.

Artificial and natural agents

One can distinguish natural agents from artificial agents. Some agents are *natural* in the sense that their existence can be explained by biological considerations; people and dogs are natural agents insofar as they exist in consequence of biological reproductive capacities – and are hence biologically *alive*. Some agents might be *artificial* in the sense that they are manufactured by intentional agents out of pre-existing materials external to the manufacturers; such agents are *artifacts*. Highly sophisticated computers might be artificial agents; they are clearly artificial and would be artificial agents if they satisfy the criteria for agency – in particular, if they are capable of instantiating intentional states that cause actions.

The distinction between natural and artificial agents is not mutually exclusive and hence should not be thought to preclude an artificial agent that is biologically alive. An example of an agent that is both artificial and natural would be certain kinds of clone. If we could manufacture living DNA out of preexisting non-genetic materials, then the resulting organism would be both artificial and biologically alive. If sufficiently complex to constitute an agent, then it would be an agent that was artificial but nonetheless alive. As a conceptual matter, something can be both artificial and biologically alive and can therefore be both an artificial and natural agent.

Nor are the concepts of artificial and natural agencies jointly exhaustive. There might be agents that are neither artificial nor natural as I have defined these notions. If, for example, an all-perfect personal God, as conceived by classical theism, created the natural universe, then God is an agent but neither an artificial nor natural one. Only an agent could *create* a universe, but God is neither biologically alive nor, according to classical theism, manufactured or created by another agent.

The concept of *moral* agency

According to the standard view, the concept of moral agency is ultimately a normative notion that picks out the class of beings whose behavior is subject to moral requirements. The idea is that, as a conceptual matter, the behavior of a moral agent is governed by moral standards, while the behavior of something that is not a moral agent is not governed by moral standards. As such, moral agents have moral obligations, while beings that are not moral agents do not have moral obligations. Adult human beings are, for example, typically thought to be moral agents and have moral obligations, while cats and dogs are not thought to be moral agents or have moral obligations.

The concept of moral agency should be distinguished from that of moral patiency. Whereas a moral agent is something that has duties or obligations, a moral patient is something *owed* at least one duty or obligation. Moral agents are usually, if not always, moral patients; all adult human beings are moral patients. But there are many moral patients that are not moral agents; a newborn infant is a moral patient but not a moral agent – though it will, other things being equal, become a moral agent.

On the standard view, the idea of moral agency (but not the idea of moral patiency) is conceptually associated with the idea of being accountable for one's behavior. To say that one's behavior is governed by moral standards and hence that one has moral duties or moral obligations is to say that one's behavior should be guided by and hence evaluated under those standards. Something subject to moral standards is *accountable* (or morally responsible) for its behavior under those standards.[3]

[3] There are potentially two distinct ideas here: (1) it is rational to hold moral agents accountable for their behavior; and (2) it is just to hold moral agents accountable for their behavior. While (2) presumably implies (1), it is not the case that (1) implies (2); while it is reasonable to think that moral standards figure into a determination of what is rational, they are not the only standards of rationality – and there might be other considerations (perhaps prudential in character) that imply the rationality of holding someone accountable. Nothing much turns on this distinction.

Although only an agent can be a moral agent, the converse is not true. The idea of moral agency is conceptually associated with the idea of being accountable for one's behavior. Dogs are agents, but not moral agents because they are not subject to moral governance and hence not morally accountable or their actions.

To hold something accountable is to respond to the being's behavior by giving her what her behavior deserves – and what a behavior deserves is a substantive moral matter. Behaviors that violate a moral obligation deserve (and perhaps require) blame, censure, or punishment. Behaviors that go beyond the call of duty (i.e., a so-called "supererogatory" act) in the sense that the agent has sacrificed important interests of her own in order to produce a great moral good that the agent was not required to produce deserve praise. Behaviors that satisfy one's obligations deserve neither praise nor censure of some kind; ordinarily, one does not deserve praise, for example, for not violating the obligation to refrain from violence.

The notion of desert, which underlies the notion of moral accountability, is a purely backward-looking notion. What one deserves is not directly concerned with changing or reinforcing one's behavior so as to ensure that one behaves properly in the future; regardless of whether one can change someone who is culpable for committing a murderer[4] by censuring him, he deserves censure. To put it in somewhat metaphorical terms, desert is concerned with maintaining the balance of justice. When someone commits a bad act, the balance of justice is disturbed by his act and can be restored, if at all, only by an appropriate act of censure or punishment.[5] When someone performs a supererogatory act, she is owed a debt of gratitude, praise or recognition; until that debt is discharged, the balance of justice remains disturbed.

These are uncontroversial conceptual claims (i.e., claims about the content of the concept) in the literature and comprise what I have been calling the standard view. As *Routledge Encyclopedia of Encyclopedia* explains the notion, "[m]oral agents are those agents expected to meet the demands of morality."[6] According to *Stanford Encyclopedia of Philosophy*, "a moral agent [is] one who qualifies generally as an agent open to responsibility ascriptions."[7] According to *Internet Encyclopedia of Philosophy*, moral agents "can be held accountable for their actions – justly praised or blamed, deservedly punished or rewarded."[8]

It is important to realize that the quotations here are not being offered as evidence that the standard view is correct; an encyclopedia reference in philosophy cannot bear that weight. Rather, they are being offered to corroborate that what I have described as the standard view is, in fact, the standard view. That this view is reproduced in encyclopedia articles without question is sufficient to show this. Again, my project here is to trace out the implications of the standard views of agency and moral agency to see what they tell us about whether consciousness is a necessary feature of agency and moral agency.

In closing this section, I should point out that these claims are conceptual in the same sense the claim that a bachelor is unmarried is conceptual: it is true in virtue of the core conventions for using the relevant terms – and will remain true for as long as those conventions are practiced.

Necessary and sufficient conditions for moral agency

The issue of which conditions are necessary and sufficient for something to qualify as a moral agent is a different issue than the issue of identifying the content of the concept. Whereas an analysis of the content of the concept must begin with the core conventions people follow in using the term, an analysis of the capacities something must have to be appropriately held accountable for its behavior is a substantive meta-ethical issue, and not a linguistic or conceptual issue.

It is generally thought there are two capacities that are necessary and jointly sufficient for moral agency. The first capacity is not well understood: the capacity to freely choose one's acts.[9] While the concept of free will remains deeply contested among compatibilist

[4] People who are insane or severely cognitively disabled, on the standard account, are not culpable and hence do not deserve censure or punishment.

[5] I say "if at all" here because an act of censure cannot erase the bad act. Punishing a murderer, for example, cannot bring her victim back to life. In such cases, it seems not possible to fully restore the balance of justice. In other cases, an act of compensation, together with censure, might be enough.

[6] Vinit Haksar, "Moral Agents," *Routledge Encyclopedia of Philosophy*.

[7] Andrew Eshelman, "Moral Responsibility," *Stanford Encyclopedia of Philosophy*.

[8] Garrath Matthews, "Responsibility," *Internet Encyclopedia of Philosophy* (James Fieser, ed.); available at http://www.iep.utm.edu/r/responsi.htm#SH2a.

[9] Not surprisingly, this entails that only agents are moral agents. Agents are distinguished from non-agents in that agents initiate responses to the world that count as *acts*. Only something that is capable of acting counts as an agent and only something that is capable of acting is capable of acting freely.

and libertarian conceptions, there are a few things that can be said about it that are uncontroversial among both compatibilist and libertarian conceptions. On either account, one must, for example, be the direct cause of one's behavior in order to be characterized as freely choosing that behavior; something whose behavior is directly caused by something other than itself has not freely chosen its behavior. If, for example, A injects B with a drug that makes B so uncontrollably angry that B is helpless to resist it, then B has not freely chosen his or her behavior.

This should not be taken to deny that external influences are relevant with respect to the acts of moral agents. It might be, for example, that human beings come pre-programmed into the world with desires and emotional reactions that condition one's moral views. If this is correct, it does not follow that we are not moral agents and should not be thought to rule out the possibility of artificial moral agents programmed by other persons. All that is being claimed here is that it is a necessary condition for being free and hence a moral agent that one is the direct cause of one's behavior in the sense that its behavior is not directly compelled by something external to it.

Moreover, the relevant cause of a *moral* agent's behavior must have something to do with a decision. Consider a dog, for example, trained to respond to someone wearing red by attacking that person. Although the dog might be the direct cause of its behavior in the sense that *its* mental states produce the behavior, it has not freely chosen its behavior because dogs do not make decisions in the relevant sense. In contrast, the choices that cause a person's behavior are sometimes related to some sort of deliberative process in which the pros and cons of the various options are considered and weighed. It is the ability to ground choice in this deliberative process, instead of being caused by instincts, that partly warrants characterizing the behavior as free.

This should not be taken to mean that all free choices result from deliberation. Most people, including myself, make many decisions during the course of the day without anything resembling a process of deliberation. My choice to have a cup of coffee this morning is no less a decision or free choice because it was not preceded by a deliberation of any kind. We frequently make spontaneous decisions, based on desires, gut-feelings, or previous deliberations. The claim here is that it is not a necessary condition for an act to be free that the decision to perform that act be the outcome of some deliberative process; rather, the claim is that free acts are the results of decisions – and only a thing capable of deliberating can make a *decision*. The capacity for deliberation is thus a necessary condition for free will and hence for moral agency.

Thus, the idea that moral agents are free presupposes that they are rational. Regardless of whether one's deliberations are caused or cause one's behavior, one can deliberate only to the extent that one is capable of reasoning – and this is the hallmark of rationality. Something that acts wholly on the basis of random considerations is neither making decisions, deliberating nor acting rationally (assuming that she has not rationally decided that it is good to make decisions on such a basis). Someone who acts on the basis of some unthinking compulsion is not making decisions, deliberating, acting rationally, or freely choosing her behaviors. Insofar as one must reason to deliberate, one must have the capacity to reason and hence be rational to deliberate.

The second capacity necessary for moral agency is also related to rationality. As traditionally expressed, the capacity is "knowing the difference between right and wrong"; someone who does not know the difference between right and wrong is not a moral agent and not appropriately censured for her behaviors. This is, of course, why we do not punish people with severe cognitive disabilities like a psychotic condition that interferes with the ability to understand the moral character of her behavior.

As traditionally described, however, the condition is too strong because it seems to suggest moral infallibility as a necessary condition for moral agency – and no one is morally infallible. Knowledge, as a conceptual matter, requires justified true belief. But it is not clear that any fallible human being *knows* which acts are right and which acts are wrong; this would require one to have some sort of generally reliable methodology for determining what is right and what is wrong – and no fallible human being can claim such a methodology. In any event, this much is certainly clear: many (if not most) adult human beings, notwithstanding their own views to the contrary, do not *always* know which acts are right and which are wrong.

About the most that we can confidently say about moral agents is that they have the ability to engage in something fairly characterized as moral reasoning. This ability may be more or less developed. But anyone who is justly or rationally held accountable for her behavior must have the potential to engage in something that is reliable, much of the time, in identifying the requirements of morality. The idea that a being should conform her behavior to moral requirements presupposes that she has the ability to do so; and this requires not only that she have free will, but also that she has the potential to correctly

identify moral requirements (even if she frequently fails to do so). At the very least, it requires people to correctly identify core requirements – such as is stated by the principle that it is wrong to kill innocent persons for no reason.

Moral reasoning requires a number of capacities. First, and most obviously, it requires a minimally adequate understanding of moral concepts like "good," "bad," "obligatory," "wrong," and "permissible" and thus requires the capacity to form and use concepts. Second, it requires an ability to grasp at least those moral principles that we take to be basic – like the idea that it is wrong to intentionally cause harm to human beings unless they have done some sort of wrong that would warrant it (which might very well be a principle that is universally accepted across cultures). Third, it requires the ability to identify the facts that make one rule relevant and another irrelevant. For example, one must be able to see that pointing a loaded gun at a person's head and pulling the trigger implicates such rules. Finally, it requires the ability to correctly apply these rules to certain paradigm situations that constitute the meaning of the rule. Someone who has the requisite ability will be able to determine that setting fire to a child is morally prohibited by the rule governing murder.[10]

The conditions for moral agency can thus be summarized as follows: for all X, X is a moral agent if and only if X is (1) an agent having the capacities for (2) making free choices, (3) deliberating about what one ought to do, and (4) understanding and applying moral rules correctly in paradigm cases. As far as I can tell, these conditions, though somewhat underdeveloped in the sense that the underlying concepts are themselves in need of a fully adequate conceptual analysis, are both necessary and sufficient for moral agency.

Consciousness as implicitly necessary for moral agency

Although what I have called the standard account of moral agency does not explicitly contain any reference to consciousness, it is reasonable to think that each of the necessary capacities presuppose the capacity for consciousness. The idea of accountability, central to the standard account of moral agency, is sensibly attributed only to conscious beings. That is to say, the standard account of moral agency, I will argue, applies only to conscious beings – although this may not be true of non-standard accounts.

There are a number of reasons for this. First, it is a conceptual truth on the standard account that an action is the result of some intentional state – and intentional states are *mental* states. While this is not intended to rule out the claim that mental states are brain states and nothing else, only a being that has something fairly characterized as a conscious mental state is also fairly characterized as having intentional states like volitions – regardless of what the ultimate analysis of a mental state turns out to me. It is a conceptual truth, then, that agents have mental states and that some of these mental states explain the distinguishing feature of agents – namely the production of doings that count as actions.

Second, Jaegwon Kim argues that if we lacked some sort of access to those mental states that constitute reasons, then we would lack a first-person self-conscious perspective that seems necessary for agency.[11] It cannot, for example, be the *external* presence of a stop sign that directly causes a performance that counts as an action; the cause must have something to do with a reason that is internal – like a belief about the risks or consequences of running a stop sign and a desire to them. If I don't have some sort of access to something that would count as a reason for doing X, doing X is utterly arbitrary – akin to a random production by a device lacking a first-person perspective. Although Kim does not explicitly claim that the access must be conscious, it is quite natural to think that it must be. Reasons are *grasped* – and this is a conscious process. While grasping a reason need not entail an ability to articulate it, an agent must have some understanding of why she is doing X. If our ordinary intuitions are correct, even a dog has something resembling conscious access to the fact that she eats *because* she is hungry or *because* what is offered is tasty.

Third, as a substantive matter of practical rationality, it makes no sense to praise or censure something that lacks conscious mental states – no matter how otherwise sophisticated its computational abilities might be. Praise, reward, censure, and punishment are rational responses only to beings capable of

[10] It is worth noting that Luciano Floridi and Jeff Sanders agree that moral accountability presupposes free will and consciousness. As to free will, they assert that: "if the agent failed to interact properly with the environment, for example, because it actually lacked sufficient information or had no choice, we should not hold an agent morally responsible for an action it has committed because this would be morally unfair." See Floridi and Sanders (2001, p. 18). However, they believe moral agency does not necessarily involve moral accountability.

[11] Jaegwon Kim, Reasons and the First Person. In J. Bransen and S. Cuypers, editors, *Human Action, Deliberation, and Causation*. Kluwer, 1998. I am indebted to Rebekah Rice for pointing this out to me.

experiencing conscious states like pride and shame. As Floridi and Sanders put this plausible point, "[i]t can be immediately conceded that it would be ridiculous to praise or blame an AA [i.e., artificial agent] for its behaviour or charge it with a moral accusation. You do not scold your webbot, that is obvious" (2001, p. 17).

The reason is that it is conceptually impossible to reward or punish something that is not conscious. As a conceptual matter, it is essential to punishment that is reasonably contrived to produce an unpleasant mental state. You cannot punish someone who loves marshmallows, as conceptual matter, by giving them marshmallows; if it doesn't hurt, it is not punishment, as a matter of definition – and *hurt* is something only a conscious being can experience. While the justification for inflicting punitive discomfort might be to rehabilitate the offender or deter others, something must be reasonably calculated to cause some discomfort to count as punishment; if it isn't calculated to hurt in some way, then it isn't punishment. Similarly, a reward is something that is reasonably calculated to produce a pleasurable mental state; if it isn't calculated to feel good in some way, then it isn't a reward. Only conscious beings can have pleasant and unpleasant mental states.

Each of the substantive capacities needed for moral agency, on the standard account, also seem to imply the capacity for consciousness. It is hard to make sense of the idea of a non-conscious thing freely choosing anything. It is reasonable to think that there are only two possible explanations for the behavior of any non-conscious thing: its behavior will either be (1) purely random in the sense of being arbitrary and lacking any causal antecedents or (2) fully determined (and explainable) in terms of the mechanistic interactions of either mereological simples or higher-order but equally mechanistic interactions that emerge from higher order structures composed of mereological simples. It is not implausible to think that novel properties that transcend explanation in terms of causal interactions of atomic constituents emerge from sufficiently complex biological systems.

Indeed, the very concept of deliberation presupposes the capacity for conscious reasoning. All animals have some problem-solving capacities, but only human beings can solve those problems by means of a manipulation of concepts that are *understood*. But only a conscious being can decide what to do on the basis of abstract reasoning with concepts. Unconscious computers and non-rational sentient beings solve problems, a capacity associated with rationality, but do not do so by means of consciously reasoning with symbols. As a conceptual matter, only something that is conscious can deliberate – though the converse is clearly not true; higher animals are arguably conscious but cannot deliberate.

We might, of course, be wrong about this; but if so, the mistake will be in thinking that we freely choose our behavior. It might be that our conscious deliberations play no role in explaining our acts and that our behavior can be fully explained entirely in terms of the mechanistic interactions of ontological simples. Our sense that we decide how we will act would be, in that case, mistaken; our behavior would be as mechanistically determined as the behavior of any other material thing in the universe – though the causal explanation for any piece of human behavior will be quite complicated.

This does not mean that a behavior is freely chosen only if preceded by some self-conscious assessment of reasons that are themselves articulated in a language. To repeat an important point, very few acts are preceded by a conscious process of reasoning; most of what we do during the day is done without much, if any, conscious thought. My decision this morning to make two cups of coffee instead of three was not preceded by any conscious process of reasoning. But it seems no less free to me because I did not have to think about it. Beings that can freely choose their behavior by consciously deliberating about it can sometimes freely choose behavior without consciously deliberating about it.

Nevertheless, it seems to be a necessary condition for something to freely choose its behavior that it be capable of conscious deliberation. If there are beings in the universe with free will, then they will certainly be conscious and capable of consciously deciding what to do.

The same is true of the capacity for moral understanding: it is a necessary condition for something to know, believe, think, or understand that it has conscious mental states. Believing, as a conceptual matter, involves a disposition to assent to *P* when one considers the content of *P*; assenting and considering are conscious acts. Similarly, thinking, as a conceptual matter, involves a process of conscious reasoning. While it may turn out that thinking can be explained entirely in terms of some sort of computational process, thinking and computation are analytically distinct processes; an ordinary calculator can compute, but it cannot think. Terms like "know," "believe," "think," and "understand" are intentional terms that apply only to conscious beings.

This is a point that emerges indirectly from the debate about John Searle's famous "Chinese Room" argument that conscious states cannot be fully

explained in computational terms.[12] As is well known, Searle asks us to suppose we are locked in a room, and given a rule book in English for responding in Chinese to incoming Chinese symbols; in effect, the rule book maps Chinese sentences to other Chinese sentences that are appropriate responses. Searle argues that neither you nor the system for responding to Chinese inputs that contains you "understands" Chinese. Searle believes that the situation is exactly the same with a computer; as he makes the argument:

> The point of the story is this: by virtue of implementing a formal computer program from the point of view of an outside observer, you behave exactly as if you understood Chinese, but all the same you don't understand a word of Chinese. But if going through the appropriate computer program for understanding Chinese is not enough to give you and understanding of Chinese, then it is not enough to give any other digital computer an understanding of Chinese. And again, the reason for this can be stated quite simply. If you don't understand Chinese, then no other computer could understand Chinese because no digital computer, just by virtue of running a program, has anything that you don't have. All that the computer has, as you have, is a formal program for manipulating uninterpreted Chinese symbols. To repeat, a computer has a syntax, but no semantics. The whole point of the parable of the Chinese room is to remind us of a fact that we knew all along. Understanding a language, or indeed, having mental states at all, involves more than just having a bunch of formal symbols. It involves having an interpretation, or a meaning attached to those symbols. And a digital computer, as defined, cannot have more than just formal symbols because the operation of the computer, as I said earlier, is defined in terms of its ability to implement programs. And these programs are purely formally specifiable – that is they have no semantic content.

It is true, of course, that Searle's argument remains controversial to this day, but no one disputes the conceptual presupposition that only conscious beings can fairly be characterized as "understanding" a language. The continuing dispute is about whether consciousness can be fully explained in terms of sufficiently powerful computing hardware running the right sort of software. Proponents of this view believe that the fact that a functioning brain is contained in a living organism is irrelevant with respect to explaining why it is conscious; an isomorphic processing system made entirely of non-organic materials that runs similar software would be conscious – regardless of whether it is fairly characterized as "biologically alive." If so, then it is capable of understanding, believing, knowing and thinking. Thus, the dispute is about whether consciousness can be fully explained in terms of computational processes, and not about whether non-conscious beings can know, believe, think, or understand. Nearly all sides agree that such terms apply only to conscious beings.

Either way, it seems clear that only conscious beings can be moral agents. While consciousness, of course, is not a sufficient condition for moral agency (as there are many conscious beings, like cats, that are neither free nor rational), it is a necessary condition for being a moral agent. Nothing that isn't capable of conscious mental states is a moral agent accountable for its behavior.[13]

None of this should be taken to deny that conscious beings sometimes act in cohort or that these collective acts are rightly subject to moral

[12] As Searle elsewhere describes the view, "The brain just happens to be one of an indefinitely large number of different kinds of hardware computers that could sustain the programs which make up human intelligence. On this view, any physical system whatever that had the right program with the right inputs and outputs would have a mind in exactly the same sense that you and I have minds. So, for example, if you made a computer out of old beer cans powered by windmills; if it had the right program, it would have to be a mind. And the point is not that for all we know it might have thoughts and feelings, but rather that it must have thoughts and feelings, because that is all there is to having thoughts and feelings: implementing the right program."

[13] Floridi and Sanders argue that the idea that moral agency presupposes consciousness is problematic: "the [view that only beings with intentional states are moral agents] presupposes the availability of some sort of privileged access (a God's eye perspective from without or some sort of Cartesian internal intuition from within) to the agent's mental or intentional states that, although possible in theory, cannot be easily guaranteed in practice" (16). The problem with this view is that it does not engage the standard account if intended to do so. On the standard view, it is not the idea of a moral agency that presupposes that we can determine which beings are conscious and which beings are not; it is rather the ability to reliably determine which beings are moral agents and which beings are not that presupposes that we can reliably determine which beings are conscious and which beings are not. If moral agency presupposes consciousness, then we cannot be justified in characterizing a being as a moral agent unless we are justified in characterizing the being as being conscious.

evaluation. As a moral and legal matter, we frequently have occasion to evaluate acts of corporate bodies, like governments and business entities. The law includes a variety of principles, for example, that make it possible to hold business corporations liable under civil and criminal law.

Strictly speaking, corporate entities are not moral agents for a basic reason. A corporate entity is a *set* of objects, which includes conscious moral agents accountable for their behavior, but also includes, at the very least, legal instruments like a certificate of incorporation and bylaws. The problem here is that a set is an *abstract* object and as such incapable of *doing* anything that would count as an "act." Sets (as opposed to a representation of a set on a piece of paper) are no more capable of acting than numbers (as opposed to representations of numbers); they have nothing that would count as "state," internal or otherwise, that is capable of changing – a necessary precondition for being able to act. Sets are not, strictly speaking, moral agents because they are not *agents* at all.

The acts that we attribute to corporations are really acts of individual directors, officers, and employees acting in coordinated ways. Officers sign a contract on behalf of the organization, and new obligations are created that are backed by certain assets also attributed to the corporation. Officers decide to release a product and instruct various parties to behave in certain ways that have the effect of releasing the product into the stream of commerce. Though we attribute these acts to the corporate entity for purposes of legal liability, corporate entities, *qua* abstract objects, do not act; corporate officers, employees, etc. do.

Indeed, the law acknowledges as much, characterizing a corporate person as a "legal fiction." The justification for the fiction of treating corporations as agents is to encourage productive behavior by allowing persons who make decisions on behalf of the corporation to shield their personal assets from civil liability – at least in the case of acts that are reasonably done within the scope of the corporation's charter. If the assets of, for example, individual directors were exposed to liability for bad business decisions, people would be much less likely to serve as business directors.

Our moral practices are somewhat different and less dependent upon fictional assertions of agency to corporate entities. Most people rightly seek to attribute moral fault for corporate misdeeds to those persons who are most fairly characterized as responsible for them. It is clear, for example, that we cannot incarcerate a corporation for concealing debts to artificially inflate shareholder value, but we can – and do – incarcerate individual officers for their participation in schemes to conceal debts. We do not say Enron was bad; we say that the people running Enron were. And we would make this distinction even if every person on Enron's payroll were behaving badly.

Consciousness as implicitly necessary for agency

It turns out that the capacity for consciousness seems to be presupposed by the simpler notion of agency itself. As will be recalled, the concept of agency can be expressed as follows: X is an agent if and only if X can instantiate intentional mental states capable of directly causing a performance; here it is important to remember that intentional states include beliefs, desires, intentions, and volitions (or the relevant neurophysiological correlates). In any event, on the received view, doing α is an action if and only if α is caused by an intentional state (and is hence performed by an agent).

The problem here is that the very notion of agency presupposes the idea that the actions of an agent are caused by some sort of mental state – and mental states are conscious. While a few psychoanalytic theorists have floated the idea of unconscious mental states, this is, strictly speaking, incoherent. What, as a conceptual matter, distinguishes mental from non-mental states is, among other things, the former are privately observable by introspection, while non-mental states are publicly observable by third parties by processes that require a different mental state – namely perception, the object of which are non-mental objects. The capacity to introspect and to observe privately themselves presuppose consciousness; so if mental states are characterized by the ability to be privately observed by the subject by introspection, it follows that they are *conscious* mental states; one cannot introspect or observe what is not available to consciousness.

Similarly, the notion of an intentional state, as it is traditionally conceived, also seems to presuppose consciousness. The very first sentence of the entry on intentionality in *Stanford Encyclopedia of Philosophy* asserts "Intentionality is the power of *minds* to be about, to represent, or to stand for, things, properties and states of affairs." Similarly, the very first sentence of the entry in the *Routledge Encyclopedia of Philosophy* states that "intentionality is the *mind's* capacity to direct itself on things." By definition, minds are conscious. Thus, if the standard accounts of agency and its cognate inten-

tionality are correct, the very notion of agency itself presupposes consciousness in the sense that only a conscious being can be an agent.

Artificial agents

None of this, of course, should be taken to suggest that artificial ICTs cannot be agents or moral agents accountable for their behavior on the standard account. Rather, it is to claim that an artificial ICT can be an agent only if conscious on the standard account and that an artificial ICT can be a *moral agent* only if it is an agent with the capacities to choose its actions "freely" and understand the basic concepts and requirements of morality, capacities that also presuppose consciousness.

It is clear that an artificial agent would have to be a remarkably sophisticated piece of technology to be a moral agent. It seems clear that a great deal of processing power would be needed to enable an artificial ICT to be able to (in some relevant sense) "process" moral standards. Artificial free will presents different challenges: it is not entirely clear what sorts of technologies would have to be developed in order to enable an artificial entity to make "free" choices – in part, because it is not entirely clear in what sense *our* choices are free. Free will poses tremendous philosophical difficulties that would have to be worked out before the technology can be worked out; if we don't know what free will *is*, we are not going to be able to model it technologically.

Determining whether an artificial agent is conscious involves even greater difficulties. First, philosophers of mind disagree about whether it is even possible for an artificial ICT (I suppose we are an example of a natural ICT) to be conscious. Some philosophers believe that only beings that are biologically alive are conscious, while others believe that any entity with a brain that is as complex as ours will produce consciousness regardless of the materials of which that brain is composed.

Second, even if it should turn out that we can show conclusively that it is possible for artificial ICTs to be conscious, there are potentially insuperable epistemic difficulties in determining whether or not any particular ICT is conscious. It is worth remembering that philosophers have yet to solve even the problem of justifying the belief that other human beings than ourselves are conscious ("the problem of other minds"). Since we have direct access to only our own consciousness, knowledge of other minds would have to be indirect through some sort of argument by analogy.[14] Taking that strategy and applying it to other kinds of things, like animals and artificial ICTs, weakens it considerably because the closeness of the resemblance between us and another type of being diminishes the fewer properties that type of thing shares with human beings. Indeed, it is not at clear at this point how we could even begin to determine that a machine is conscious. The epistemological difficulties associated with trying to determine whether a machine is a moral agent are well beyond us at this point.

Even so, we might be morally obligated to treat certain sophisticated ICTs as if they are moral agents without being justified in thinking they are and hence without being able to rule out the possibility that they are not. According to the problem of other minds, I am not epistemically justified in believing that there are any other conscious minds in the world than my own; while I might try to infer as much by a behavioral and physiological analogy, this analogy is really an induction that is based on the observation of one case – namely, my own case (and each person can be sure that she is conscious). But the fact that we are not justified in thinking that other people have minds doesn't entail that we ought not to treat them as moral agents accountable for their behavior. If something walks, talks, and behaves enough like me, I might not be justified in thinking that it has a mind, but I surely have an obligation, if our ordinary reactions regarding other people are correct, to treat them as if they are moral agents. The above analysis is not only agnostic with respect to the issue of whether conscious computers are possible, but also with respect to the issue of whether computers that seem conscious (in the same way that other people seem conscious) is sufficient to give rise to a moral obligation to treat them as if they are moral agents and hence morally accountable for their behavior.

Conclusions

In this essay, I have described and given the justifications for the standard accounts of the concepts of agency, moral agency, moral responsibility, as well as

[14] But philosophers of mind have shown that such analogical similarities may not be sufficient to justify thinking someone is conscious. In essence, someone who infers that X is conscious based on X's similarity to him is illegitimately generalizing on the strength of just one observed case – one's own. Again, I can directly observe the consciousness of only one being, myself; and in no other context is an inductive argument sufficiently grounded in one observed case. The further from our own case some entity is, the more difficult it is for us to be justified in thinking it is conscious.

described the standard meta-ethical analysis of the substantive conditions a thing must satisfy to be accountable for its behavior. I have argued further that the conditions for agency and moral agency, together with the moral conditions for accountability, all presuppose consciousness. I have concluded that while there are difficult epistemic issues involved in determining whether an artificial ICT is conscious and a moral agent, it is a necessary condition for an artificial ICT to be a moral agent that it is conscious. I have not, however, drawn any conclusions about how artificial agents that appear conscious (though the appearance is not enough to warrant believing they are conscious) should be treated.

References

K. Coleman. Computing and Moral Responsibility. *Stanford Encyclopedia of Philosophy*. Available at http://plato.stanford.edu/entries/computing-responsibility/, 2004.

A. Eshleman. Moral Responsibility. *Stanford Encyclopedia of Philosophy*. Available at http://plato.stanford.edu/entries/moral-responsibility/, 2001.

L. Floridi. Information Ethics: On the Philosophical Foundation of Computer Ethics. *Ethics and Information Technology*, 1(1): 37–56, 1999.

L. Floridi and J. Sanders. Artificial Evil and the Foundation of Computer Ethics. *Ethics and Information Technology*, 3(1): 56–66, 2001.

K.E. Himma. What is a Problem for All is a Problem for None: Substance Dualism, Physicalism, and the Mind-body Problem. *American Philosophical Quarterly*, 42(2): 81–92, 2005.

D. Johnson. Computer Systems: Moral Entities but not Moral Agents. *Ethics and Information Technology*, 8(4): 195–204, 2006.

F.W.J. Keulartz et al. Pragmatism in Progress. *Techne: Journal of the Society for Philosophy and Technology*, 7(3): 38–49, 2004.

J. Kim. Reasons and the First Person. In J. Bransen and S. Cuypers, editors, *Human Action, Deliberation, and Causation*, pp. 67–87. Kluwer Publishers, Dordrecht, 1998.

B. Latour. On Technical Mediation – Philosophy, Sociology, Genealogy. *Common Knowledge*, 3: 29–64, 1994.

K. Miller and D. Larson. Angels and Artifacts: Moral Agents in the Age of Computers and Networks. *Journal of Information, Communication & Ethics in Society*, 3(3): 113, 2005.

J. Moor. Reason, Relativity, and Responsibility in Computer Ethics. *Computers and Society*, 28(1): 14–21, 1998.