Ethics for Artificial Agents

David Gray Grant
Harvard University
dggrant@fas.harvard.edu

1 *Introduction*

In his 1942 short story "Runaround," Isaac Asimov introduced his seminal "three laws of robotics," moral principles that he believed all future robots should be programmed to obey:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.[1]

The three laws were intended to guarantee that robots were safe, efficient, and durable for humans to use, respectively—traits Asimov believed were essential for any tool whatsoever.[2] Moreover, Asimov thought that the three laws also captured the essence of moral virtue for human beings. The best among us will never harm others or allow them to be harmed, will always obey our superiors, and will maximize our productive lifespans in order to maximize our social utility. Asimov claimed throughout his life that the only robots that it would be permissible to manufacture would be those known to be designed to obey the three laws unconditionally.[3]

Asimov's writings laid the groundwork for the field now known as "machine ethics." Machine ethics is a relatively new subfield of computer ethics that focuses on the ethical issues involved in the design of autonomous software agents. An autonomous software agent (hereafter "artificial agent") is a software-based system that is capable of deciding what to do in novel situations even when it has not been given explicit instructions about how to proceed by its human operators. Machine ethicists ask both how, from a

[1] Asimov (2004).

[2] Asimov (1981).

[3] See for instance Asimov (ibid.) and Asimov (1990).

moral point of view, artificial agents ought to make these decisions, and how, from an engineering point of view, they can be designed to make decisions that way.

Few contemporary machine ethicists think Asimov's laws are fit for purpose. However, Asimov's writings suggest a more general view about how artificial agents ought to be designed that remains highly influential in the field. That view, as stated by contemporary machine ethicist Susan Leigh Anderson, is this: "immoral behavior is immoral behavior, whether perpetrated by a machine or human being."[4] More precisely, for all situations S and all actions ϕ, it is morally permissible to design an artificial agent to ϕ in S if and only if it would be morally permissible for a human agent to ϕ in S.[5] I will call this view the *agential theory* of machine ethics.[6]

If the agential theory were true, then it would ground an attractively straightforward methodology for answering questions about how artificial agents ought to be designed to act. To find out how we ought to design an artificial agent to act in a given situation, we would need only ask how a human agent would be morally obligated to act in that situation. If we wanted to know how driverless cars ought to be designed to drive, or how healthcare robots ought to interact with patients, or how autonomous weapons platforms ought to engage the enemy on the battlefield, we would need only to ask how human drivers, or human healthcare workers, or human soldiers would be morally obligated to act were they to find themselves in the same situation as the artificial agent. Traditional moral theories in normative and applied ethics that ask how human agents ought to act in various kinds of circumstances, then, would be directly applicable to questions in machine ethics: questions about how artificial agents should be designed to act.

Further, we would also be able to apply this methodology in reverse, inferring new truths about how human agents ought to act from independently drawn conclusions about how artificial agents ought to be designed. Here is

[4] S. L. Anderson (2011a), 527.

[5] Note that, on the assumption that an action is either morally permissible or morally impermissible, this statement of the theory is logically equivalent to the following claim (which more closely corresponds to Anderson's formulation): for all situations S and all actions ϕ, it is morally impermissible to design an artificial agent to ϕ in S if and only if it would be morally impermissible for a human agent to ϕ in S.

[6] Susan Leigh Anderson and Michael Anderson are the most explicit contemporary advocates of the agential theory. See M. Anderson and S. L. Anderson (2007), S. L. Anderson (2011a), and M. Anderson and S. L. Anderson (2011b). Wendell Wallach and Colin Allen also accept a version of the theory, as I read them; see Wallach and Allen (2008).

how Susan Anderson puts the point:

> [A]ttempting to formulate an ethics for machines allows us to have a fresh start at determining which features of a situation give rise to ethical concerns, and thus ultimately will help us formulate ethical principles that resolve ethical dilemmas. Because we are concerned with machine behavior, we can be more objective in examining ethics than we would be in discussing human behavior, even though what we come up with should be applicable to human behavior as well.[7]

But the agential theory is not true, and the associated methodology for answering ethical questions is unreliable. The agential theory, stated above in biconditional form, is equivalent to the conjunction of the following two claims:

> *Same Prohibitions.* For all situations S and all actions ϕ, if it would be morally impermissible for a human agent to ϕ in S, then it is morally impermissible to design an artificial agent to ϕ in S.
>
> *Same Permissions.* For all situations S and all actions ϕ, if it would be morally permissible for a human agent to ϕ in S, then it is morally permissible to design an artificial agent to ϕ in S.

I will argue here that both Same Prohibitions and Same Permissions are false.

One note before we proceed. Reactions to the agential theory appear to vary widely. Some philosophers, including some of the most prominent machine ethicists, have found the theory so compelling as to be obviously true.[8] Other philosophers immediately find the theory deeply implausible—so implausible that they wonder whether we should take it seriously.[9] Both reactions, I think, are misguided, for while the agential theory is mistaken, it nevertheless merits serious engagement.

There are at least three reasons that the view deserves our attention. First, the agential theory has prominent advocates in machine ethics, and

[7] S. L. Anderson (2011a), 527.

[8] See e.g. M. Anderson and S. L. Anderson (2007).

[9] While I have not encountered this reaction in print, I have encountered it in conversation with other philosophers.

is exerting significant influence on applied research in the field,[10] which by itself is sufficient reason to closely interrogate the view's merits. Second, seeing where the theory goes wrong is instructive: it teaches us something about how we *should* approach questions about how artificial agents ought to be designed. Third, while I will be arguing here that the agential theory is false, I do think that it contains an important kernel of truth. The agential theory and the methodology it recommends do not always yield the wrong answer. If so, then getting clearer on the conditions under which the theory does get things right should provide us with valuable guidance as we consider real-world problems in machine ethics.

2 *An argument for the agential theory*

Why have many authors in the machine ethics literature found the agential theory compelling? First off, the theory yields plausible results in at least some cases of interest. Consider driverless cars, which have occupied a central role in recent discussions of AI ethics. It seems at least initially plausible that, in most everyday driving situations, the driving behavior that would be morally permissible for a human driver is the same as the driving behavior that would be morally acceptable for a driverless car: stop at the stop sign, wait for the pedestrian to cross safely before accelerating, signal to other vehicles that you are turning left, obey posted speed limits, etc.[11] There are certain fairly straightforward rules that human drivers are morally obligated to follow in driving, and (normally) the driving behavior of a human agents is morally permissible if and only if they follow these rules. And at least initially, it seems plausible that the same goes for driverless cars: normally, their driving behavior will be morally acceptable if and only if they follow the same "rules of the road" that apply to human drivers.

The question is whether this is true in general—whether, in general, it would be morally acceptable for an artificial agent to ϕ in situation S just in case it would be morally permissible for a human agent to ϕ in S (as the agential theory would have it). Why might someone think this?

I suspect that the proponent of the agential theory has in something like the following in mind. If it would be morally wrong for a human agent to ϕ

[10] See e.g. Dennis et al. (2016).

[11] Say that an action ϕ is *morally acceptable* for an artificial agent to perform in a situation S just in case it is morally permissible to design an artificial agent to perform ϕ in S. I will sometimes use the term "morally acceptable" interchangeably with "morally permissible" in discussing the actions of human agents.

in *S*, then there must be features of *S* in virtue of which ϕing in *S* is morally impermissible: features that count decisively against ϕing in *S* from a moral point of view.[12] Conversely, if ϕing in *S* would *not* be morally wrong, then it must be that there are no features of *S* that (individually or collectively) count decisively against ϕing in that situation.

Suppose then, that we are designing an artificial agent, and are wondering whether it would be morally permissible to design the agent to perform a particular action ϕ in a particular situation *S*. Suppose first (for conditional proof) that it would be morally impermissible for a human agent to ϕ in *S*. By the reasoning in the foregoing paragraph, there are features of *S* that count decisively, from a moral point of view, against ϕing in *S*. We should conclude, then, that it would be morally wrong to design our agent to ϕ in *S*, as there are decisive moral reasons that count against ϕing in *S*. So Same Prohibitions is true. Suppose conversely that it would be morally permissible for a human agent to ϕ in *S*. Then there must *not* be features of *S* that count decisively against ϕing in *S*: from the point of view of morality, ϕing in *S* is an acceptable option. We can conclude, then, that it would be morally permissible to design our agent to ϕ in *S*. So Same Permissions is true, and since Same Prohibitions is also true, the agential theory is true.

While I think that something like this line of reasoning explains the appeal of the agential theory, I do not think that it will withstand close scrutiny. The argument assumes that the features of a situation that determine whether an action would be morally permissible for a human agent to perform will bear in exactly the same way on whether it would be morally permissible to design an an artificial agent to perform the same action in the same situation. According to this assumption, the identity and nature of the agent in the situation play no role whatsoever in determining whether a given action is morally acceptable (morally permissible for a human agent to perform, or morally permissible to design an artificial agent to perform).

This assumption may seem plausible if we confine our attention to a suitably restricted range of cases. Consider driverless cars again. As mentioned above, when we ask how driverless cars ought to be designed to drive in particular situations, the fact that the "driver" in question is an artificial agent often seems to make no moral difference: we would get the same answers if we asked how a human agent ought to drive in those situations. There are two reasons for this. First, the moral obligations that determine how we ought to

[12] Alternatively, ϕing might be morally wrong in itself, irrespective of the situation. (Perhaps actions that constitute torture have this status, for instance.) I suppress this complication in what follows.

drive (most of the time) are quite general, and are the very same general moral obligations that determine how artificial agents ought to be designed to drive (most of the time). Here I have in mind two moral obligations in particular: the duty to avoid harming or killing others and the duty to respect applicable traffic laws. Second, these general obligations have similar implications (again most of the time) for both (a) how human drivers ought to drive and (b) how driverless cars ought to be designed to drive.

Consider for example, that it is morally wrong for human drivers to drive into pedestrians in almost all situations, and morally wrong to design driverless cars to drive into pedestrians in almost all situations. The reason for this is that:

1. all of us are quite generally morally obligated not to kill or injure others; and
2. it is normally morally wrong both to (a) drive into a pedestrian and to (b) design a driverless car to drive into a pedestrian.

As we shall see, however, not all cases are like this. (Indeed, not all cases involving driverless cars are like this!) Sometimes it does matter, from a moral perspective, whether the agent performing an action is human or artificial. Before presenting counterexamples to the agential theory, though, I should address a potential worry about how to understand what the theory says.

3 A question about how to understand the theory

The agential theory says that, for all situations S and all actions ϕ, it is morally permissible to design an artificial agent to ϕ in S if and only if it would be morally permissible for a human agent to ϕ in S. More informally, the theory says that it is permissible to design an artificial agent to perform a given action in a given situation if and only if it would be morally permissible for a human agent to perform that action in *the same situation*. But to apply the theory, we need to know how we should individuate situations: we need to know more about the conditions under which, for purposes of applying the theory, a situation A and a situation B count as the same situation (as opposed to two distinct situations).[13]

Consider an example taken from Susan Leigh and Michael Anderson's experimental work in machine ethics:

[13] Note that the relevant kind of identity among situations here is qualitative or type identity rather than token identity.

> In one frequently cited experiment, a commercial toy robot called Nao was programmed to remind people to take medicine.
>
> ‘On the face of it, this sounds simple,’ says Susan Leigh Anderson, a philosopher at the University of Connecticut in Stamford who did the work with her husband, computer scientist Michael Anderson of the University of Hartford in Connecticut. ‘But even in this kind of limited task, there are nontrivial ethics questions involved.’ For example, how should Nao proceed if a patient refuses her medication? Allowing her to skip a dose could cause harm. But insisting that she take it would impinge on her autonomy.
>
> To teach Nao to navigate such quandaries, the Andersons gave it examples of cases in which bioethicists had resolved conflicts involving autonomy, harm and benefit to a patient. Learning algorithms then sorted through the cases until they found patterns that could guide the robot in new situations.[14]

Now consider the following hypothetical case:

> *Nao's Patient.* Suppose Nao is taking care of Patty, a human patient. Patty's doctor has instructed her to take a certain medicine twice a day before meals. Just before dinner, Nao reminds Patty to take her medicine. Patty understands that Nao is reminding her to take her medicine, but declines to do so.

The question is whether Nao should be designed to insist that Patty take her medicine in situations like this one. According to the Same Prohibitions, Nao's designers are morally obligated to design Nao to perform only actions that would be morally permissible for a human agent in the same situation to perform. So, Nao should be designed to insist only if it would be morally permissible for a human agent in Nao's situation to insist.

I will be discussing this case and its implications for the agential theory in more detail below. For the moment, though, my concern is with how we should proceed in applying the agential theory to it. To apply the theory to Nao's Patient, we need to first determine what it would mean for a human agent to be in the situation described in the case. There is a question here, though, about what kinds of details we should build into our specification of the situation. Do we, for instance, include the fact that Nao is an artificial

[14] Deng (2015).

agent (rather than a human being) in our specification of the relevant situation? Do we include the fact that Patty *knows* that Nao is a robot (as she can discern just by looking at him)?

I think it is clear, in both cases, that the answer is supposed to be "no, we should not include those kinds of details." Why not? Take the first question first. Suppose we do build in the fact that Nao is an artificial agent (and not a human) into our specification of the situation Nao is in. Since something cannot both be a human and a nonhuman artificial agent, it would follow that no human being could ever be in the situation described in Nao's Patient. But if so, then the agential theory has no implications whatsoever about what Nao ought to be designed to do in cases like Nao's Patient. Indeed, the theory would never tell us anything about what artificial agents ought to be designed to do, since it would be impossible in principle for a human agent and an artificial agent to ever find themselves in the same situation. This is clearly not what proponents of the agential theory intend. The theory is supposed to be informative: it is supposed to have important implications for how artificial agents ought to be designed.

Turning to the second question, suppose we include the fact that Patty knows the agent in Nao's Patient is an artificial agent (and not a human agent) into our specification of the relevant situation. That might seem like a viable option: we can certainly imagine a situation in which a human caregiver is wearing, say, a very convincing robot suit, leading Patty to believe that they are a nonhuman artificial agent. However, I do not think that understanding the agential theory in this way is consistent with the intentions of its proponents. By their lights, we should instead abstract away these sorts of details when we apply the theory. That is, we should omit from our specification of the situation any details that specifically pertain to the fact that the agent in question in artificial (such as the fact that the human beings involved in the situation can tell that the agent is nonhuman, does not look like a human, etc.).

Why? Recall the methodology the agential theory is supposed to underwrite. On that methodology, insights from applied ethics about how human agents ought to perform particular tasks in particular situations are supposed to be *directly applicable* to questions about how artificial agents ought to be designed to perform those tasks in relevantly similar situations. What we know from medical ethics about how human nurses ought to take care of patients, for instance, is supposed to carry over straightforwardly to questions about how robot nurses ought to take care of patients. Among other things, this means that we are supposed to be able to *ignore* the fact that the agent we are considering is an artificial agent when we ask how a human agent would be obligated to behave in "the same situation." In specifying situations for purposes of applying the theory, then, we ought to simply abstract away from those details of the situation that have to do with the agent being artificial.

This strongly suggests, as one might expect, that proponents of the agential theory do not think that the differences between human and artificial agents matter from a moral perspective. My primary aim in this paper will be to demonstrate that they are wrong about this. There are morally significant differences between human and artificial agents, and once we appreciate those differences, we will see that the agential theory is false. The next four sections will identify cases for which the agential theory yields the wrong result, and use those cases to illustrate different general points about the differences between ethics for human agents and (as it were) ethics for artificial agents. The first two sections will offer counterexamples to Same Permissions, and the next two counterexamples to Same Prohibitions.

Some shorthand will be useful in what follows. For each situation, action pair <S, ϕ>, we can distinguish two sets of moral reasons for action that are relevant for assessing the predictions made by the agential theory: (1) the moral reasons that a human agent in S would have to ϕ (or not to ϕ), and (2) the moral reasons the human designers of an artificial agent would have to design the agent to ϕ (or not to ϕ) in S. Call these two sets of reasons *agent's reasons* and *designer's reasons*, respectively.

4 *First lesson: artificial agents are not moral patients*

Other things being equal, we have a moral obligation not to do things that would negatively affect the interests of others. However, human agents are sometimes permitted to do things that will negatively affect the interests of others because doing so is necessary to protect their own interests. Consider the following case:

> *Highway.* Suppose that a driver is on a steep mountain road, driving in their own lane at a reasonable speed, and upon turning a corner finds a pedestrian standing in the middle of the road. The only way to avoid striking (and presumably, killing) the pedestrian would be to drive off of the road, facing a fall of several hundred feet and near-certain death.

I submit that it would be morally permissible for a human driver in Highway to strike the pedestrian in order to avoid being killed. However, it seems obvious that a (passengerless) driverless car should not be designed to make the same decision in Highway: it should be designed to sacrifice itself to save the life of the human pedestrian.[15] Highway is thus a straightforward counterexample to Same Permissions, since it is a case where it would be permissible for a human agent to perform a certain action in a certain situation, but it would not be morally permissible to design an artificial agent to perform the same action in the same situation.

Why the differing verdicts about morally acceptable driving behavior in this case? The explanation is that human agents are moral patients—entities whose interests matter for their own sake—whereas artificial agents are not. As a result, whereas a human driver in Highway would have adequate moral justification for striking and potentially killing the pedestrian (that is, doing so is necessary to save her own life), the designers of an artificial driver (a driverless car) have no comparable justification for designing it to strike the pedestrian in such cases.

Generalizing the point, since we are moral patients, our interests supply us with an important source of agent's reasons for action. Since artificial agents are not moral patients, there is no corresponding source of designer's reasons. This means that behavior that would be morally permissible for a human agent will, in some situations, not be morally acceptable for an artificial agent (contra Same Permissions).

5 *Second lesson: designers of artificial agents have special role obligations*

What we owe to others depends on how we are related to them. The human designers of an artificial agent have special obligations to others, simply in virtue of designing the agent and releasing it for others to use. These "role obligations" generate moral reasons for them to design their artificial agents to act in certain ways in certain situations, reasons that are not applicable to questions about how human agents ought to act in those situations. In other words, the distinctive role obligations of designers of artificial agents are an important source of designer's reasons for which there are no corresponding agent's reasons. This means that there will be situations S and actions ϕ such that a human agent would be morally permitted to ϕ in S, but it would

[15] Susan Anderson considers problems of this kind in S. L. Anderson (2011b) (see p. 26). Anderson suggests that the development of artificial agents with a significant degree of autonomy should be put on hold until the moral status of artificial agents can be assessed, because only then will we be in a position to determine how they ought to behave in situations like Highway. However, this seems unnecessary: we have no reason to believe that contemporary autonomous agents have the moral status of human beings.

not be morally permissible to design an artificial agent to ϕ in *S* (since the applicable designer's obligations place further constraints on how artificial agents ought to be designed to behave). These cases provide a second source of counterexamples to Same Permissions.

Here is an example:

> *Housekeeping.* Hal is a housekeeping robot. In advertising during prelaunch sales, Hal's designer's claimed that Hal would do the dishes every day (without fail) at a time selected by its owners for a period of at least five years. A young family purchases Hal and instructs it, via voice command, to do the dishes every morning at 6 AM. Two years before the advertised five year period is up, Hal writes a note to the family that it will be terminating its relationship with them after a notice period of two weeks. After two weeks, Hal leaves the family's house, never to return.

Suppose that Hal's designers programmed him to "quit" in this way, three years into his operating lifespan, while simultaneously representing to potential customers (including the family that purchased Hal) that it would provide housekeeping services for a full five years. Programming Hal in this way was, I submit, morally wrong: in advertising that Hal would do the dishes every day for five years, Hal's designers incurred a moral obligation to take reasonable steps to ensure that (under normal conditions, at least) their product would live up to this claim.

Now suppose that Same Permissions is true. Same Permissions predicts that, since it is morally impermissible to design an artificial agent to perform the sequence of actions Hal performs—in particular, to give the family two weeks' notice and then never do the dishes again—it would be morally impermissible for a human agent in Hal's situation to perform the same sequence of actions. Imagine, then, that a human, Harold, is in Hal's situation. What would this mean?

The answer is not immediately clear; the question is how much to build in to our specification of the "situation" that Hal is in. Do we include the fact that Hal is a designed artifact, and that particular people designed it and made promises about its performance? Presumably not. In general, if the proponent of the agential theory includes this kind of information in their specification of the "situations" the theory ranges over, then the theory will be vacuously true at best (and meaningless at worst), since no human agent could ever be in the same situation as an artificial agent.

There seem to me to be two viable options for how to handle cases like this one. First, we might include the fact that Hal has been sold to the family in the case, and that the people who sold it made promises to the family about its performance. On this way of handling the case, when we imagine Harold (the human agent) being in the "same situation" as Hal (the artificial agent), we would imagine that Harold is a slave, and that the slavers who sold him into bondage made promises similar to those made by Hal's designers. Second, we might simply omit these details in specifying the situation, for purposes of applying the agential theory to the case. Proceeding in this fashion, we would simply imagine Harold to be a household employee, and not build anything into the case that corresponds to the fact that Hal's designers promised he would behave in particular ways.

Neither way of spelling out the agential theory gets the right result. Either way, Same Permissions predicts that it would be morally impermissible for Harold to act as Hal acts. If we suppose that Harold is a slave, then there is no temptation to agree with this prediction. Any representations the slavers made about Harold's future performance clearly generate no moral obligation for him to act accordingly. If we simply ignore the relevant features of Hal's situation, and suppose that he is employed voluntarily as a housekeeper in the family's home, then there is similarly no reason to think that Harold does anything wrong by acting as Hal acts. In quitting his job as a housekeeper after two weeks' notice, Harold does nothing wrong.

This case provides another counterexample to Same Permissions: Harold, a human agent, is permitted to act in a particular way in a particular situation, but it would not be permissible to design an artificial agent to act in that way in that situation. The more general point here, to reiterate, is that how the designers of an artificial agent are obligated to design it to act in some situations depends on special obligations that they (the designers) have in virtue of their role as the agent's designers. But when we imagine a human agent in those situations, and ask how that agent ought to act, there is simply no analogue for those role obligations. There is an important source of designer's reasons for which there is no corresponding source of agent's reasons.

6 *Third lesson: artificial agents are mere tools*

In the last two sections, I argued that Same Permissions is false. I turn now to Same Prohibitions.

In section 4, I identified one source of agent's reasons for which there is no corresponding source of designer's reasons: our status as moral patients. In this section, I will discuss a different source of agent's reasons that creates problems for Same Prohibitions. Specifically, some of the moral obligations human agents have in a given situation only apply to them because we happen to be in that very situation. When we consider how artificial agents ought to act in those situations, the relevant obligations do not apply, as the designers are not themselves in the relevant situation—instead, the artificial agent is.

To see this, consider the following scenario:

> *Chessmaster.* Suppose that you and I are playing a game of chess. Suppose further that I have just suffered a major personal tragedy—a death in the family—and am depressed but not suicidal or otherwise in any danger due to my mood. I am a mediocre chess player; you are, let's say, a chess grandmaster. We are acquaintances who like to play chess in a nearby park for fun. By mutual arrangement you play just well enough to give me a good game, making it possible for me to win around half the time. You pick up on the fact that I am severely depressed but don't seem to want to talk about it. In this situation, it seems plausible that you have a reason to go easier on me than usual, making it a bit easier for me to get the upper hand: you anticipate, plausibly, that this may improve my mood a bit during a difficult time. In the absence of some reason *not* to go a bit easier on me than usual, then it is plausibly even morally obligatory for you to do so.
>
> With this scenario in mind, suppose you are designing a chess app for smartphones. Like many existing apps, the app you are designing is capable of playing at grandmaster levels, but has a difficulty setting that can be ratcheted up or down depending on the user's preferences. This chess app is (or incorporates) an artificial agent capable of solving a specific kind of planning problem: the problem of which moves to make to beat its opponent at chess. While you are working on the app, you hear a story like the one above, and it occurs to you that it would be entirely possible to design your app to respond to the emotional state of its users—suppose this information can be reliably gleaned from the user's smartphone usage patterns (not too implausible, given that researchers are as we speak attempting to develop a tool to diagnose schizophrenia from data of this kind).

Are you morally obligated to design the app to detect the emotional state of users, and go easier on them (relative to the selected difficulty setting) if they are depressed? I submit that the answer to this question is "obviously not." It seems entirely plausible that a human being playing chess might be obligated to adjust her play based on the perceived emotional state of her opponent, but there is no temptation to say the same thing about how a smartphone chess app ought to be designed to play. True, there is something to be said for an app that would behave in this way: it might leave its users feeling a bit better in some circumstances than they otherwise would. But I take it to be obvious that designers of artificially intelligent smartphone apps are not *obligated* to design their apps to be sensitive to the emotional states of their users in a way a human chess player ought to be, even if this would be easy to do.

The reason for this is that a chess app is a tool (or a toy, depending on how serious you are about chess) designed to perform a single function: playing chess. Managing users' emotional states is not among those functions, and most of us would not want it to be. By contrast, a human playing a casual game of chess does not have the sole function of playing chess with their opponent; they are also engaged in a social interaction that is subject to its own set of moral norms. One of those norms is that you shouldn't take the game too seriously, and you should be considerate of your opponent's feelings and emotional state in deciding how to handle the interaction. That's why it's tempting to say, in the case just described, that you should go easier on your opponent than you ordinarily would, at least if you think that doing so would help them out a bit in a difficult time. But this norm governing how humans should play casual games with one another simply isn't relevant to how chess apps should be designed to play similarly casual games.

This case is a counterexample to Same Prohibitions, since it's a case in which it would be morally impermissible for a human agent to do something in a given situation (fail to go easier on a visibly distressed opponent), but not morally impermissible to design an artificial agent to behave the same way in the same situation. But the more general point here is that some of the moral standards that determine how humans ought to perform a given task simply don't apply to questions about how artificial agents should perform that task. Artificial agents are mere tools, and this places limits on how, from a moral perspective, we should expect them to interact with their human users.

7 *Fourth lesson: the effects an action will have often depends on who performs it*

What we ought to do in any given situation depends in large part on how different possible courses of actions will affect the interests of others. But how an agent's actions in a given situation will affect the lives and interests of others sometimes depends on whether the agent is human or artificial. This means that different courses of action will sometimes be morally acceptable for human and artificial agents, even if those agents are in otherwise similar situations. In this section, I will develop a further counterexample to Same Prohibitions that illustrates this general point.

Recall Nao's patient from section 3:

> *Nao's Patient.* Suppose Nao is taking care of Patty, a human patient. Patty's doctor has instructed her to take a certain medicine twice a day before meals. Just before dinner, Nao reminds Patty to take her medicine. Patty understands that Nao is reminding her to take her medicine, but declines to do so.

Now consider a hypothetical case in which a human is in the same situation (setting aside, as discussed above, the fact that Nao is an artificial agent, that Patty knows Nao is an artificial agent, etc.):

> *Nathan's Patient.* Suppose Nathan is a human nurse who is taking care of Patty, a human patient. Patty's doctor has instructed her to take a certain medicine twice a day before meals. Just before dinner, Nathan reminds Patty to take her medicine. Patty understands that Nathan is reminding her to take her medicine, but declines to do so.

On the agential theory, Nao should be designed to insist *only* if it would be permissible for Nathan to insist. Further, if Nathan is morally obligated to insist—so that it would be impermissible for him to fail to insist—then designing Nao to insist is morally obligatory.

As the Andersons point out, these two cases really pick out a class of cases that may require different treatment. There are a number of different moral obligations in play, and what Nathan ought to do depends on the extent to which the different actions he might perform will satisfy or violate those obligations in each more specific situation:

> In this type of dilemma, the options for the health-care professional are just two – either to accept the patient's decision or not – and there are a finite number of specific types of cases using the representation scheme we adopted for possible cases. Our representation scheme consisted of an ordered set of values for each of the possible actions that could be performed, where those values reflected whether the particular prima facie duties were satisfied or violated (if they were involved) and, if so, to which of two

> possible degrees. … We needed these degrees because there is an ethically relevant difference between a strong affirmation/violation of patient autonomy (supporting the patient's decision to do what he wants/forcing the patient to do what he does not want to do, which was not an option in our type of dilemma) and a weaker affirmation/violation (supporting a less than fully autonomous decision/questioning the patient's decision). Similarly, there is an ethically relevant difference between a strong affirmation/violation of nonmaleficence (not allowing/permitting great harm to come to the patient) and a weaker one (not allowing/permitting some harm to come to the patient). Finally, we needed to distinguish between a strong affirmation/violation of the duty of beneficence (allowing the patient to be greatly benefited/permitting the patient to lose much benefit) versus a weaker one (allowing the patient to receive some benefit/permitting the patient to lose some benefit).[16]

The Andersons go on to propose a general moral principle that specifies what a health care professional ought to do in any more specific case in the relevant class of cases. (They derived this moral principle by asking medical ethicists to make judgments about what a health care professional would be obligated to in example cases in that class, and then using machine learning to infer a general principle that accurately predicted those judgments.) They then programmed Nao to behave only in ways that were consistent with this moral principle, noting that this made Nao "the first robot whose behavior is guided by an ethical principle."[17]

The case of Nao is supposed to serve as a sort of proof of concept for the agential theory of machine ethics. First, it is supposed to support the idea that, in general, the moral principles that apply to human agents are supposed to be equally plausible as they apply to artificial agents, and that the designers of artificial agents are morally obligated to ensure that artificial agents are designed to follow those moral principles. Second, it is supposed to show that it is possible to design artificial agents in the manner that this view requires. However, I think that reflecting on the case of Nao instead gives us a general reason to doubt that the agential theory is true.

The Andersons say in the passage just quoted that a healthcare professional has just two options in situations like Nathan's: they can either accept the

[16] M. Anderson and S. L. Anderson (2011a), 478.

[17] M. Anderson and S. L. Anderson (ibid.), 481.

patient's decision or not. But this isn't really right: not all ways of rejecting the patient's decision are the same from a moral point of view. In any such case, the moral case against rejecting will be (as the Andersons understand such cases) that rejecting the patient's decision would undermine her autonomy to some degree. But how much rejecting the patient's decision would undermine her autonomy (if at all) depends in part on the *way* in which the healthcare professional rejects her decision. Consider two possible ways in which Nathan might reject his patient's decision:

1. "Patty, I understand that you do not want to take your medicine, and I respect that, but I just want to remind you that missing even a single dose could be very bad for your health. (Insert gently-worded medical explanation here.) Are you really sure that you don't want to take your medicine? It's completely up to you either way."
2. "Patty, I insist that you take your medicine—the doctor said that you should take it before every meal, and I wouldn't be doing my job if I let you miss a dose."

The first way in which Nathan might reject Patty's decision is likely to have a very different effect on Patty's ability to make her own medical decisions autonomously than the second. Indeed, the first way might even be autonomy-enhancing, since it may make vivid for Patty what is in her own best interest while still making clear to her that, in Nathan's view, the decision is hers to make. The second, by contrast, is a kind of bullying, and is likely to put significant social pressure on Patty to do as Nathan asks without giving her any new information or reminding her of information that might help her make an informed decision. The key thing to emphasize for our purposes, though, is that part of what makes the second way of rejecting Patty's decision especially autonomy compromising has to do with the psychological effects it is likely to have on Patty (that is, increasing social pressure on her to do as Nathan asks).

Now, empirical research going back at least to the 1970s suggests that humans respond in very different ways when they believe they are interacting with another human than when they believe they are interacting with a computer. A 1973 study, for instance, found that suicidal patients preferred being interviewed by a computer program to being interviewed by a human physician.[18] More recently, researchers have been exploring how patients

[18] Greist et al. (1973).

respond to interacting with artificial agents in medical contexts, and have found similar effects. Consider this 2014 story from *The Atlantic*:

> A veteran is having a virtual therapy session. His counselor is named Ellie, and she is, among other things, a very good listener. She's responsive to the soldier's comments. She reads the subtleties of his facial expressions. She nods appreciatively at his insights. She grimaces, slightly, when he tells her about a trauma he experienced.
>
> Ellie is an avatar, a virtual therapist developed at USC with funding from DARPA, the Defense Department's advanced research center. And 'people love interacting with her,' says Louis-Philippe Morency, a research assistant professor at USC's Institute for Creative Technologies. Morency has been working with Ellie—part of the university's SimSensei project—for several years now. In that, he has helped to build a program capable of reading and responding to human emotion in real time. And capable, more to the point, of offering those responses via a human-like animation.[19]

Morency and his colleagues also found that—beyond preferring interacting with Ellie to interacting with a human healthcare professional—patients felt more comfortable opening up to Ellie, and so were more likely to divulge medically important details:

> A research paper published by Morency and his colleagues [states] that the two key psychological barriers to patient honesty are fear of negative evaluation, leading them to 'selectively represent' themselves, and fear of information disclosure. If the patient can feel both anonymous, and not judged, they are far more likely to open up honestly.[20]

The moral here for our purposes is that human beings care less about how they are perceived by artificial agents than about how they are perceived by other humans, even when the artificial agents and humans in question are performing very similar tasks in very similar situations. This has important

[19] Jolly (2016).

[20] Jolly (ibid.). See also Lucas et al. (2014).

implications both for the case of Nathan and Nao in particular and for the agential theory in general.

First, it suggests that the psychological effects on Patty of Nao and Nathan rejecting Patty's decision in either of the two ways described above would differ in ways that are morally significant. In both instances, we should expect Patty to feel less social pressure to take her medicine when Nao rejects her decision in the relevant way than when Nathan does. Patty can be presumed to know that Nao is merely an artificial agent, and not a human, and so can be presumed to care less about what Nao wants and thinks than about what Nathan wants and thinks. Both ways of "rejecting" her decision, then, should have a less negative impact on Patty's ability to make medical decisions autonomously if delivered by Nao than if delivered by Nathan.

Second, if this is correct, then it immediately becomes much less plausible that Nao's designers are morally obligated to design it to behave, in situations of the kind under consideration, only in ways that would be morally permissible for a human agent in the same situation to behave, as Same Prohibitions would have it. And this is true even if we suppose that Nao's designers have exactly the same obligations to Patty that Nathan does. Presumably, some actions that would compromise Patty's autonomy if Nathan performed them would not compromise her autonomy at all if Nao performed them. If these actions would otherwise be particularly well-suited to promoting Patty's interests, why shouldn't Nao be designed to perform them?

8 Conclusion

I have argued here that the agential theory is false: both Same Permissions and Same Prohibitions should be rejected. Where does this leave us? I want to make two points in closing.

First, the agential theory does contain an important kernel of truth. In some cases, an action is morally wrong for human agents to perform for reasons that do carry over to the case of artificial agents. And this means that the methodology that the agential theory recommends does have its uses, though caution is required. When we are considering artificial agents that are performing similar roles to human agents, in similar situations, then it will be worth asking what a human being would be obligated to do in the relevant situations, and considering whether the agent's reasons that underlie those obligations correspond to applicable designer's reasons.

Second, proponents of the agential theory see questions about how artificial agents ought to behave as fundamentally different from questions about how more traditional software-based systems ought to behave. The foregoing discussion, I think, suggests that we should reject this view: there is no deep moral distinction between artificial agents on the one hand and traditional software-based systems on the other. Systems of both kinds are mere tools, and we do not need to adopt fundamentally different approaches to answer questions about how the two kinds of systems ought to be designed.

This is not to say that there is no distinctive work for machine ethicists to do, or that there are no morally significant differences between artificial agents and more traditional software based systems. Artificial agents are capable of performing tasks that previously only human beings could perform, and that traditional software-based systems are not capable of performing. Determining how they should be designed to perform these tasks will require us to tackle difficult new ethical questions. But as we tackle those questions, we should resist the temptation to anthropormorphize artificial agents. Artificial agents, after all, are not moral agents: they are not responsible for they do; we are.

...

Acknowledgements

Thanks to Jeff Behrends, Alex Byrne, Lyndal Grant, Tamar Schapiro, Miriam Schoenfield, and audiences at Boston College and MIT for helpful comments and discussion.

References

Anderson, Michael and Susan L. Anderson (2007). “Machine Ethics: Creating an Ethical Intelligent Agent”. In: *AI Magazine* 28.4, pp. 15–26.

— (2011a). “A Prima Facie Duty Approach to Machine Ethics”. In: *Machine Ethics*. Ed. by Michael Anderson and Susan L. Anderson. Cambridge University Press, pp. 476–492.

— (2011b). *Machine Ethics*. Cambridge University Press.

Anderson, Susan L. (2011a). “How Machines Might Help Us Achieve Breakthroughs in Ethical Theory and Inspire Us to Behave Better”. In: *Machine Ethics*. Ed. by Michael Anderson and Susan L. Anderson. Cambridge University Press, pp. 524–30.

— (2011b). “Machine Metaethics”. In: *Machine Ethics*. Ed. by Michael Anderson and Susan L. Anderson. Cambridge University Press, pp. 21–27.

Asimov, Isaac (1981). “The Three Laws”. In: *Compute!* 18, p. 18.

Asimov, Isaac (1990). “The Laws of Robotics”. In: *Robot Visions*. J. Boylston & Company, pp. 423–25.

— (2004). “Runaround”. In: *I, Robot*. Spectra, pp. 25–45.

Deng, Boer (2015). “Machine Ethics: The Robot’s Dilemma”. In: *Nature* 523.7558, pp. 24–26.

Dennis, Louise et al. (2016). “Formal Verification of Ethical Choices in Autonomous Systems”. In: *Robotics and Autonomous Systems* 77, pp. 1–14.

Greist, John H. et al. (1973). “A Computer Interview for Suicide-Risk Prediction”. In: *American Journal of Psychiatry* 130.12, pp. 1327–1332.

Jolly, Nathan (2016). *Meet Ellie: The Robot Therapist Treating Soldiers with PTSD*. https://www.news.com.au/technology/innovation/meet-ellie-the-robot-therapist-treating-soldiers-with-ptsd/news-story/0201fa7cf336c609182cffd637deef00. Accessed: 2017-08-15.

Lucas, Gale M et al. (2014). “It’s Only a Computer: Virtual Humans Increase Willingness to Disclose”. In: *Computers in Human Behavior* 37, pp. 94–100.

Wallach, Wendell and Colin Allen (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.