Bachelor's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Humanities
Department of Philosophy and Religious Studies

Christiansen, Simon

Can a machine have moral agency?

An exploration of the field of Machine Ethics and when and how to affix moral agency to machines

Bachelor's thesis in Philosophy
Supervisor: Kiran, Asle Helge
December 2022

NTNU
Norwegian University of
Science and Technology

Christiansen, Simon

Can a machine have moral agency?

An exploration of the field of Machine Ethics and when and how to affix moral agency to machines

Bachelor's thesis in Philosophy
Supervisor: Kiran, Asle Helge
December 2022

Norwegian University of Science and Technology
Faculty of Humanities
Department of Philosophy and Religious Studies

NTNU
Norwegian University of
Science and Technology

Contents

Introduction

As technology is continuing to improve rapidly, we are entrusting continuously more responsibility onto machines to be a positive impact in the world. It should be no surprise to anyone that we're therefore seeing an increasing number of ethically problematic applications of technology. Examples of this are: Self-driving cars, targeted advertising, and we have the technology to build lethal robot soldiers if we should so desire. The latter paints an unnecessarily dystopic image for the purpose of this essay. We only need look at self-driving cars to see some of the problems. Who is responsible if such a car were to do something wrong? Do we affix moral responsibility on the programmer, the owner or the car itself? It might seem a bit dangerous to not affix some responsibility onto the first two, as that would incentivize better design and regular maintenance of the car. However, is there a case to be made for allowing affixing moral agency to the car itself? Would this come with some benefits? What could be some of the downsides? How would we go about defending such an idea, and how could it possibly be implemented in practice?

In this essay, I will show that there is a need for allowing affixing moral responsibility onto machines. I will use the classical AI Ethics dilemma of the self-driving car to show how, since human top-down ethical frameworks fall short in solving the issue even for a human driver, we need something else for a self-driving car: bottom-up. By this I am referring to the self-driving car itself being a moral agent, able to make the right decision regardless of any pre-defined moral framework. This should provide it with the flexibility to tackle the wide array of situations which we would have trouble providing an answer to all of, before the car hits the road.

The essay explores Sullins' ideas about moral responsibility and criteria for affixing moral agency to a robot. More precisely, we will look at the requirements for autonomy, intentionality and responsibility. In this regard, we shall se how a lot the bias towards a machine that has moral agency is solved once we allow robots to be put to no more scrutiny than we would towards a human. This is done by showing how Sullins dismisses the need for personhood, consciousness and free will – as these are problematic philosophical concepts even for human individuals.

Later we take a quick look at Hess and his argument about non-conscious corporate moral agents. I show how I think his argument about corporations is analogous to a robot, with the core of the issue being consciousness. And where Hess draws upon Kant, I move on to Powers' article about imagining a Kantian machine. We will then see a proposal for how one could

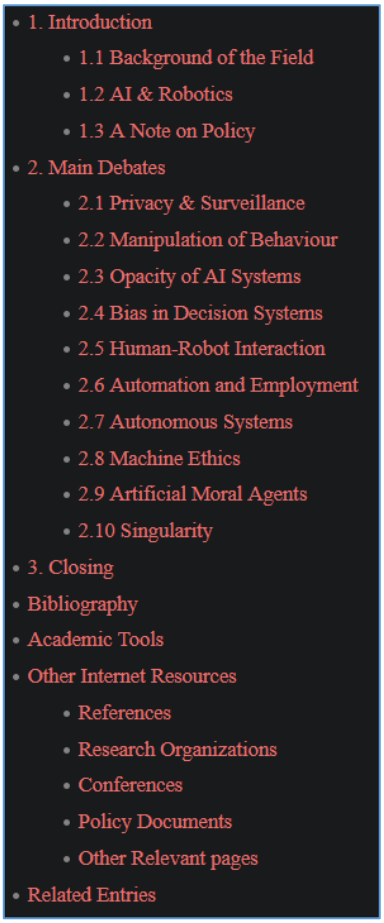
actually implement some form of self-legislative procedure within a robot, which would then remove the need for a pre-defined ethical framework to be installed in, say, a self-driving car.

We shall also finally see how the problem of consciousness for moral agency does not have to be an unrealistic feat for AI to achieve, since there are some striking similarities between how the human brain is structured and how the AI software is structured.

AI Ethics and Machine Ethics

One question that arose quite early while delving into this topic was this: For all the examples of ethical problems for AI, aren't they just as ethically challenging regardless of AI? Take the self-driving car that must choose between running over a pedestrian or driving into a wall to the detriment of those in the car, isn't this just as hard to answer for a philosopher regardless of an AI driving the car or not? Do you know what the right thing to do would be? Does it make a difference if you have a kid in the car? As you can see, this dilemma isn't really about AI. It's about ethics in general. What supposedly makes it AI Ethics is that the one faced with the dilemma would be a software program. But it seems to me like the only meaningful difference this makes, is a mere technical challenge: How do you code ethics? This is trivial to most philosophers, compared to: What ethics should be coded? From what I've read about the field of AI Ethics, it is not interested in the nitty gritty details of software development. So, if this is the only meaningful difference between an AI driving the car, and a human being driving the car, what then is left for AI Ethics other than "mere" ethics?

Vincent C. Müller wrote an SEP article called "Ethics of Artificial Intelligence and Robotics". There we find the following breakdown of sections:



(Müller, 2021)

And in the introduction, he makes the distinction between "Ethical issues that arise with AI systems as *objects*, i.e., tools made and used by humans [section 2.1 through 2.7]" (Müller, 2021), and "AI systems as *subjects* [section 2.8 and 2.9]" (Müller, 2021). It's therefore apparent that my question earlier about the distinction between AI Ethics and general ethics, is only relevant to what's referred to as the field of Machine Ethics. And in Müller's section about Machine Ethics, I could not find a definition that gave me an answer:

1. "machine ethics is concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable. (Anderson and Anderson 2007: 15)" (Müller, 2021).

   This definition doesn't satisfy, because if we know what's ethically acceptable behavior towards humans, then it is merely a question about how to program that ethic into a computer. And if we don't know what's ethically acceptable behavior towards humans, then we ought to figure that out regardless of machines.

2. “AI reasoning should be able to take into account societal values, moral and ethical considerations; weigh the respective priorities of values held by different stakeholders in various multicultural contexts; explain its reasoning; and guarantee transparency. (Dignum 2018: 1, 2)” (Müller, 2021).

   Aside from the ability to explain its reasoning and guarantee transparency, the former points seem to merely be an attempt to roughly define what ethics is. And the two latter points really belong to Müller’s section 2.3 about the opacity of AI Systems and therefore not Machine Ethics. Although I will concede that the ability to explain oneself is something we should expect from an AI-system-as-subject, and at the same time something we should make an AI-system-as-object capable of.

But I did come across this definition, which in Müller’s article was the closest I came to something else than either mere technical challenge or mere general ethics: “Some of the discussion in machine ethics makes the very substantial assumption that machines can, in some sense, be ethical agents responsible for their actions, or “autonomous moral agents” (see van Wynsberghe and Robbins 2019).” (Müller, 2021). I find the word “can” especially interesting, since it then becomes a question of whether a machine can ever be regarded as a moral agent from a philosophical standpoint, not so much whether it is easy to implement in software by an engineer.

Machine Ethics – AI systems as subjects

The introductory section of the book *Machine Ethics* (Anderson & Anderson, 2011, p. 7) starts with referencing James Moor’s four categories in which machines fall into in the context of ethics and morality:

- Normative agents (not ethically relevant)
- Ethical impact agents (not designed with ethics in mind – but has an ethical impact)
- Implicit ethical agents (designed with ethics in mind)
- Explicit ethical agents (able to make ethical decisions on their own)

Based on my previous section, the category of Ethical impact agents belongs to the broader definition of AI Ethics, and not machine ethics. An example of this could be targeted advertising, which has some ethical impacts on society, but which was seemingly not designed with ethics in

mind. Implicit ethical agents also belong to the broader field of AI Ethics, and an example of this could be self-driving cars which by themselves are not able to make ethical decisions on their own but are designed with ethics in mind. That is to say, if the engineers had not put in place some sort of ruleset based on ethics, the ethical impact of the car would be left up to chance. The question of what ruleset to implement is what I was referring to in the previous chapter. It is a question relevant to human drivers as well. But since we're having such a hard time solving that issue, might it be more convenient to attempt turning machines into moral agents themselves, and thus circumvent the need to define an ethical framework? What would happen with responsibility if we did?

Moral responsibility

Something we seem to forget when discussing the ethical challenges of self-driving cars, is that we have had this technology for many hundreds of years already. Self-driving *carts* to be precise. How is a horse carriage any different? Sure, you could steer it in a general direction and have it halt, or trot faster. This is what we would see in a self-driving car as well. The human should be able to tell it generally where to go and ask it to stop when needed. But the execution of these instructions is done by a different agent than the human. This is the same for horse carriages. Now, suppose someone runs in front of the horse, making the horse change direction and ending up running over another pedestrian. Where would we put the blame? The one who trained the horse? The owner? The one who ran in front? Or the horse? Is this analogy even comparable to that of the self-driving car?

That analogy is inspired by what John P. Sullins writes about in his article in *Machine Ethics*. He explores the idea that we can't dismiss all technology as a mere "… instrument that advances the moral interest of others" (Anderson & Anderson, 2011, p. 152). He uses a guide dog for the blind as an example, although with a focus on where to affix the moral praise, as opposed to my focus on moral blame and responsibility. The problem is the same regardless. Sullins claims both the dog trainer and the dog shares the praise.

> *The question now is whether the robot is correctly seen as just another tool or if it is something more like the technology exemplified by the guide dog. Even at the present state of robotics technology, it is not easy to see on which side of this disjunction that reality lies* (Anderson & Anderson, 2011, p. 153).

Sullins then makes the distinction between "telerobots [and] autonomous robots" (Anderson & Anderson, 2011, p. 154), stating that it is the latter category in which it seems difficult to affix

moral responsibility. He compares the software programmer to that of a parent, having a developmental impact on the robot/child which plays a role in how it behaves morally in the future, all the while moral responsibility falls mainly on the robot/child in virtue of being an autonomous moral agent. But a robot is not a person. Sullins argues that's not necessary. He refers to a theory claiming that the importance we put on personhood is the cause of many a paradox in moral theory.

> *… the way around* [these paradoxes] *is to adopt a 'mind-less morality' that evades issues like free will and intentionality, because these are all unresolved issues in the philosophy of mind that are inappropriately applied to artificial agents such as robots* (Anderson & Anderson, 2011, p. 157).

What is meant by "mind-less morality" you ask? The solution has to do with abstraction.

> *If an agent's actions are interactive and adaptive with their surroundings through state changes on programming that is still somewhat independent from the environment the agent finds itself in, then that is sufficient for the entity to have its own agency (Floridi 2004)* (Anderson & Anderson, 2011, p. 157).

The way Sullins draws on abstraction here is very similar to how David Chalmers uses abstraction when arguing for the possibility of AI consciousness when referring to "the abstract causal organization" (Chalmers, 1997, p. 247), which he calls *Functional Organization*. It is not a coincidence that the use of abstraction is similar, and in similar settings (robots). It has to do with holding robots to no more scrutiny than we would to ourselves. In the case of consciousness, we should according to Chalmers acknowledge that our own brain is made up of simplistic neurons which function could easily be replaced by a silicon chip, which is how he argues that what's necessary for consciousness is not necessarily the material the neurons are made of, but rather its function in relation to other neurons and inputs/outputs – hence "abstract causal organization" (Chalmers, 1997, p. 247). In very much the same way, Sullins and Floridi argue that we should acknowledge that our actions are in many ways determined by our past, and we end up quickly debating free will in general if we start examining moral responsibility on that note. And isn't this the core of the issue regarding robots and morality? How can we put moral responsibility onto some machine we know is programmed and predetermined? Well, if we humans aren't completely free from being predetermined, how can we expect it from robots in order to affix moral responsibility to them. By denying them that, on the mere grounds that they are programmed, we are at the same time partly denying ourselves the same thing. Therefore,

according to Sullins we should ascribe moral value and responsibility to both humans as well as robots based on the degree of which our actions are interactive and adaptive to our surroundings. That is what defines our agency.

Of course, we usually recognize that we should judge people less harshly if they did something wrong when in panic or when acting on instinct. "He was not himself", we might say. We can look at this as the persons lower-level cognition having taken over, and that tells us that we normally see someone as morally responsible based on their higher-level cognition. This is not a direct objection to Sullins here, but it raises the question of whether Sullins' theory would require us to judge people just as harshly regardless of their state of mind – which seems problematic. Or, one could argue that when in a panicked state and acting only on instinct, that one is no longer "somewhat independent from the environment the agent finds itself in". And in the case of the simplistic self-driving AI we have today, it's hard to argue that they are different. But as will be shown later, Sullins does not argue that today's AI is to be labeled moral agents. Rather that there's no bars holding them from being labeled as such in the future.

Sullins defines three dimensions to measure a robot's moral agency: Its level of autonomy, intentional behavior, and position of responsibility.

Sullins - Autonomy

On autonomy, Sullins merely requires that "… the machine is not under the direct control of any other agent or user" (Anderson & Anderson, 2011, p. 158). And if the machine works in concert with other machines or software entities, "… we simply raise the level of abstraction to that of the group and ask the same questions of the group" (Anderson & Anderson, 2011, p. 158). With "questions" as in plural, Sullins refers to the next two; intentionality and responsibility.

Sullins - Intentionality

Sullins does not use the term *intentionality* in its strongest sense, "as that is impossible to prove without argument for humans as well" (Anderson & Anderson, 2011, p. 158) – a contention which has been relevant several times in this text already. It seems as if it is only necessary for Sullins that a robot's actions are "… seemingly deliberate or calculated" (Anderson & Anderson, 2011, p. 158) in order to satisfy this criteria. He states explicitly that the actions do not have to really be intentional "… in a philosophically rigorous way" (Anderson & Anderson, 2011, p. 158). I find this a bit too convenient. One could also ask why Sullins then chose to use this term, when it is being written in a philosophical context. He does not explain this further unfortunately. But he states that "All that is needed at the level of the interaction between the

agents involved is a comparable level of personal intentionality and free will between all the agents involved" (Anderson & Anderson, 2011, p. 158). I wish Sullins was more specific by what he means with "comparable", and also explain why he now calls it "personal intentionality" when having earlier dismissed personhood as relevant for a robot's moral agency.

Sullins – Responsibility

Sullins argues a robot has moral agency if the only way to make sense of its actions is by assuming it is doing it from an understanding that it "has the duty to care for its [moral] patients" (Anderson & Anderson, 2011, p. 159). After reading this chapter, it is clear that Sullins draws the line at any actual understanding, or any actual belief residing within the robot. It only has to be "apparent" as he calls it. However, Sullins is clear about this not lessening the strength of this criteria, because although it only has to be an apparent understanding within the robot, and that his understanding does not have to come from a place of consciousness, Sullins sees this criteria as the one which will be hardest to achieve in engineering. "This would be quite a machine and not something that is currently on offer … it is going to be some time before we meet mechanical entities that we recognize as moral equals" (Anderson & Anderson, 2011, p. 159).

Sullins – Summarized

Sullins argues that by allowing robots' moral agency to be judged on the same grounds as we judge humans, and by acknowledging that issues like consciousness and free will are just as problematic for robots as it is for humans, then there seems to be no reason to dismiss the possibility of sometime developing robots with moral agency and responsibility. And Sullins gives us a framework to gauge a robots moral agency, in virtue of its autonomy, intentionality and responsibility. A theme going through Sullins' reasoning is something of a protest towards the importance of consciousness and free will. He referred to these as being "… somewhat philosophically dubious entities" (Anderson & Anderson, 2011, p. 159). It is reasonable to be skeptical of these "entities", but I won't rule out the possibility that this sentiment might show some bias coming from Sullins.

I think Sullins' framework for gauging robot moral agency is especially interesting in the way that it is compatible with gauging human moral agency. Sullins argues in such a way that he never limits his framework to robots, but is flexible enough to include both robots, humans and various kinds of animals.

Although his framework is somewhat forgiving in some cases, like the unimportance of free will, consciousness or any "ghost in the shell" at all, Sullins is clear about a robot like this being miles

away still. Especially due to the criteria for the robot to understand its responsibility. He does not mean "understanding" in its strongest sense, as in having a consciousness, but even still it requires a lot more than the software being developed today is capable of, he claims.

Is consciousness needed?

Now although Sullins argues there isn't need for personhood in order for a machine to be recognized as a moral agent, I recognize that a lack of personhood – as well as the question of free will, is one of the more pressing objections to the idea in general, which is why I decided to spend a some more time exploring this specifically. Kendy M. Hess has written a journal article titled "Does the Machine Need a Ghost? Corporate Agents as Nonconscious Kantian Moral Agents" (Hess, 2018). His paper deals mainly with corporate and other social organizations, but I was curious if his arguments would then also hold for a robot. He raises the specific question, "does a moral agent have to be phenomenally conscious?" For clarity, phenomenally conscious is a term used for consciousness in philosophy of mind, as opposed to how consciousness is defined medically. Thomas Nagel defines phenomenal consciousness as "that there is something it is like to be [an organism]" (Nagel, 1974).

Hess starts by defining what makes a corporation an agent, or what gives it agency:

> *The many elements of the structure provide multiple, overlapping reinforcements to ensure that the collective actions of the members effectively pursue the things the corporate agent desires (those things toward which the structure reliably guides members' actions) in accordance with the "picture" of the world that forms the content of corporate beliefs* (Hess, 2018, p. 70).

What he is pointing to here is the need to show that the corporate agent itself can impose its desires [albeit non-conscious] onto its members' actions. In other words, is it possible that the corporation "… could come to adopt commitments that none of its members held or desired" (Hess, 2018, p. 71)? Because if that is not the case then it would show nothing else than how the members of a corporation have moral responsibility. For the corporation itself to be a moral agent, it must be able to do both good and wrong regardless of what its, say, board members, were to decide – or refrain from deciding.

One example is what Hess calls distributed decision making. In this example, the corporation has the commitment of producing steel additives in an environmentally friendly manner. To act on this commitment, the project is divided into different departments. Some work towards reducing costs, while some improves efficiency and other parts of the process. The example then shows

how, although each department's work may seem "innocuous and rational within its own limited sphere, the new production line results in a continuing discharge that-as it continues unabated over time-pollutes a local river" (Hess, 2018, p. 72). This might then wrongly get interpreted at accepted practice, for it to then actually become accepted practice. And that's one way a corporation can end up doing something that is not the desire or commitment of its members, thereby granting it agency and moral responsibility.

Is this applicable to a robot? Suppose a robot has the overarching goal to serve its human owner. The robot sees its owner smoking and takes notice of the packaging saying that smoking kills. It also notices how the human starts coughing when smoking and concludes that this human is hurting itself. One day the human asks the robot to go buy some cigarettes, and the robot refuses. It has then gone against what was commanded. This would be a display of agency according to Hess. Now, it is admittedly hard to imagine such a robot since this would, as Sullins points out, require a substantial degree of sophistication. However, since what we're talking about here is not far from creativity in general, we need only look at the very recent advances in creative AI to start question just how far into the future a robot like this can possibly exist (Holz, 2022).

Hess then goes on to tackle whether his moral agent (a corporation) would qualify as a Kantian moral agent. But for this I'll rather look at what Thomas M. Powers wrote on that note in his "Prospects for a Kantian Machine" (Anderson & Anderson, 2011).

Implementing Kantian ethics in a machine

Until now we have discussed whether a robot needs to have consciousness, free will, and what implications regarding such a robot would have for where to affix moral responsibility. But how would we go about implementing actual moral judgement within a robot? Powers argue Kant is the obvious solution, as it provides an ethical framework already somewhat close to computing: "Rule-based ethical theories like Immanuel Kant's appear to be promising for machine ethics because they offer a computational structure for judgement" (Anderson & Anderson, 2011, p. 464). It is worth mentioning that although one of the strengths of a computer is to be able to take vast amounts of information into account, Powers reminds us that mere utilitarianism's thorny problem of never really being able to fathom all the future implications an action might have. This problem necessarily outweighs even computer's ability to calculate. It seems much more robust then to take inspiration from the deontological tradition, where some actions are deemed bad or good out of principle, logically, rather than by consequence. Powers outlines in his article three different interpretations of Kant's categorical imperative. I will focus mainly on his first

interpretation, which he labels “straightforward deductions of actions from facts” (Anderson & Anderson, 2011, p. 465).

The first principle he makes use of is naturally universalization: “Act only according to that maxim whereby you can at the same time will that it should become a universal law” (Anderson & Anderson, 2011, p. 465). He calls this a procedure for producing ethical rules, meaning this is just the principle, and that it would serve to generate some set of ethical rules. It is then of course necessary that these resulting rules are compatible with each other. So, if an action fits with the universalization principle *and* it is compatible with prior accepted ethical rules, then it is allowed into the set of ethical rules. The set of ethical rules are, according to Powers, to be taken as a list of acceptable maxims. “… an agent’s moral maxims are instances of universally quantified propositions that could serve as moral laws – that is, laws holding for any agent” (Anderson & Anderson, 2011, p. 466). So, say the robot considers an action, it would have a reason to consider this action, meaning it has an action plan. This action plan has causes and goals. And it is in virtue of these that the robot could determine through which maxim he was about to act. For example: A robot might want to sabotage for a newer robot model in order to not be replaced – so as to continue serving the household (which is one of its overarching goals). The maxim in this scenario would be something like “It is ok to sabotage others for ones owns gain”. You could argue it isn’t for its own gain if it does it for the sake of the household. But in terms of heightening its chances of being able to fulfill *its* assigned duties, it is for its own gain. Now, could this maxim be universalized? Could the robot honestly want that all other robots, and humans for that matter, sabotages each other whenever convenient? This quickly becomes a question of the robot’s ability to comprehend, and that will become relevant later. But let’s assume it concludes that it would not lead to a better world, and it would not lead to the betterment of the household, and the maxim is also slightly contradictory considering how it might be hard to sabotage for others if oneself is getting sabotaged all the time. Powers tells us the according to his theory for a Kantian machine, this maxim would be put into one of the “traditional *deontic categories* – namely, forbidden, permissible, obligatory actions” (Anderson & Anderson, 2011, p. 466). Later, if another action would translate to the same maxim, the universalization step is not necessary, as it has already been done. And although Powers doesn’t mention this himself, at least not explicitly, I believe it would be reasonable to not require every robot to have the super sophisticated ability to universalize maxims, as long as they are equipped with a database with pre-generated maxims mapped onto the correct deontic category according to the universalization step. We could expect this to be a database that could be updated over the internet. And thus, it would clash with Sullin’s definition of agency, especially regarding the

criteria for the robot to be independent. Then again, Sullin's would ask us to raise the level of abstraction to also include the mainframe hosting the database. But the requirement would still hold, that the moment this database is being written to directly by a human, the responsibility would then, at least partly, or in some cases, fall on that human.

Powers has to this point only gotten us so far. He has shown how Kantian ethics are compatible with computation on an abstract level. But an implementable program remains to be seen. He does however point out some of the main challenges developing such a software would have. One of them has to do with whether a computer can understand the difference between "I ought to do x" and "x ought to be the case". Also, and maybe most crucial, how could we develop software that would actually apply the correct "… logical connectives between circumstances, purposes, and actions; material implication (if-then) is clearly too weak" (Anderson & Anderson, 2011, p. 468). Some other issue that Powers mentions in his own summary are also relevant for any human who were to apply Kantian ethics. Like, how would one know which maxim to discard when a contradiction arise? Suppose you have only one maxim already, and a new contradictory comes to join the set. The new one shouldn't be dismissed merely because it came in second. But again, similar to other cases I've pointed to previously, Powers agrees that this is not unique to machine ethics. Once again we see an example of how the need to do work on ethics in general is relevant. But, as Powers puts it: "Perhaps work on the logic of machine ethics will clarify the human challenge as well" (Anderson & Anderson, 2011, p. 475).

Back to the self-driving car

Let's apply what we've learned so far to the example of the self-driving car that is faced with an impossible situation. Of course we shall not try to give an explicit answer to the situation, but rather explore what it would mean for this self-driving car to be an autonomous moral agent in this situation. First of all, according to Sullins, the car checks the first two conditions because "the machine is not under the direct control of any other agent or user" (Anderson & Anderson, 2011, p. 158) and its actions are "seemingly deliberate or calculated" (Anderson & Anderson, 2011, p. 158). Now as for the responsibility: A car may very well seem as if it understands its responsibility, but that is only because its actions seem deliberate, and the software developers have made it so that it is deliberately careful. But it does not seem obvious to me that any of the self-driving cars I've heard of display a sufficient apparent understanding of its responsibility to the degree that Sullins require. Sullins would not require the self-driving car to actually understand responsibility, but at least display close-to-human-level degrees of it.

I think the self-driving car holds up to Hess' requirements. As we've seen, Hess argues consciousness is not necessary for moral agency, but he requires the entity to be able to impose its commitment onto itself and nearby environment regardless of what commitments its constituents may have (like the wheels or the camera sensors). A self-driving car is necessarily always able to change its course. It is continuously revising its previous action plan in light of new information. There is in other words a brain that is making decisions at every turn, literally and figuratively.

Now, as for what kind of ethical framework engineers might have success in implementing in such a self-driving car, Powers shows promise for Kantian ethics. However, since universalizing a maxim could take anything between a millisecond and several days depending on the complexity of the universalization (some maxims might not even be possible to test for universalization), it is not something we should rely on the self-driving car to figure out as it encounters a possibly lethal split-second decision. It would be much better to generate a coherent set of universalized maxims beforehand, which the car could use to determine the deontic category of a given action (which should correspond to an already universalized maxim). That way we achieve the efficiency needed for situations like these. Of course, we do not expect the same efficiency and deontic precision from most humans. So this is one example of how machines can possibly outrank even ourselves when it comes to ethics. This all assumes Kantian ethics is even valid, which I have taken for granted for the sake of argument. It could of course be the case that Kantian ethics is not valid, in which case it might be hard to find a similar ethic which is just as computationally compatible.

Back to consciousness

As stated before, I recognize that many will find it difficult to affix moral agency to something that is not conscious, despite the arguments and theories outlined in this essay. Therefore, I quickly want to point out that the possibility of conscious machine is itself another subject in philosophy, in which there are both supporters and deniers of the idea. One famous within philosophy of mind who supports this idea, and argues quite well in favor of it, is David Chalmers. He builds upon a functionalist theory of consciousness to achieve this, in his book *The Conscious Mind: In search of a fundamental theory* (Chalmers, 1997). His theory is somewhat outdated technically, and general to all machines, but artificial intelligence is quite different than how traditional and normal software is designed. Therefore, a link between Chalmers' theory of consciousness and AI can be found in "Artificial Consciousness" (Christiansen, 2022). There you can see how what Chalmers refers to as combinatorial-state-automata – an abstraction of

biological neurons, are identical in function to the constituents of an artificial neural network used in AI. This is relevant because it once again shows how similar we humans are to machines, at least when allowing ourselves to view things with a certain level of abstraction. And through this we might recognize that we shouldn't so quickly dismiss machines as never ever being able to achieve human-level intelligence, consciousness, or moral agency.

Conclusions

In this essay I have given three different accounts on different aspects of machine ethics. We have seen how Sullins do not think consciousness or free will is necessary for moral agency, and that moral responsibility can be affixed to a robot if it is sufficiently autonomous, has apparent intentionality, and understands its overarching duty and responsibility. The latter being such a hard feat to accomplish in engineering, that we may still be far way from ever inventing such a robot, but it nonetheless being possible in principle. We have also seen how Hess argues against the need for consciousness specifically, in his argument for how non-conscious corporate entities can have moral agency. I have showed how a robot is somewhat analogous to his examples, and that it is possible to imagine a robot displaying agency according to Hess, all the while not having consciousness. And finally, we saw how Powers argues for a procedure that machines could use to generate ethical rules based on Kant and the categorical imperative. Through translating actions into corresponding maxims, which in its universalized form is put into one of the deontic categories, an action's moral value can thus be judged without there being consciousness in the robot.

And if it is the case that the reader holds consciousness to be necessary for moral agency, as even Kant could argue in his later formalizations of his ethics. Then we need only investigate the field of philosophy of mind to see that there are some promising theories arguing for the possibility of conscious machines, or at least conscious AI – due to their inner workings being quite similar to humans'.

References

Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine Ethics*. Cambridge University Press.

Chalmers, D. J. (1997). *The Conscious Mind: In search of a fundamental theory*. Oxford University Press.

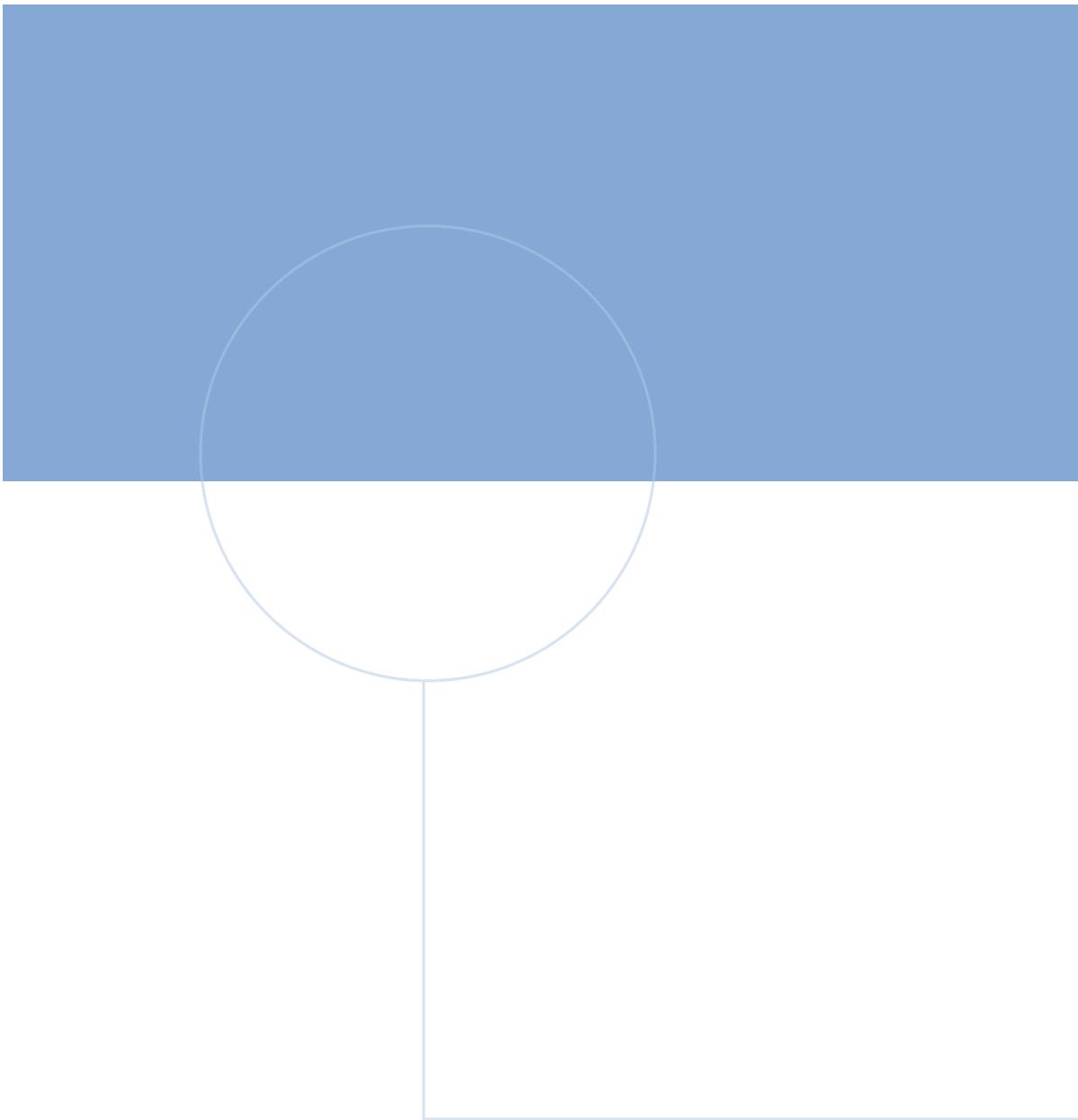
Christiansen, S. (2022). Artificial Consciousness. *Begrep Tidsskrift for filosofi*(5), 27-45.

Hess, K. M. (2018). Does the Machine Need a Ghost? Corporate Agents as Nonconscious Kantian Moral Agents. *Journal of the American Philosophical Association*, *4*(1), 67-86. https://doi.org/https://doi.org/10.1017/apa.2018.10

Holz, D. (2022). *MidJourney Community Showcase*. MidJourney. https://midjourney.com/showcase/recent/

Müller, V. C. (2021). *Ethics of Artificial Intelligence and Robotics*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435-450. https://doi.org/https://doi.org/10.2307/2183914



NTNU

Norwegian University of
Science and Technology