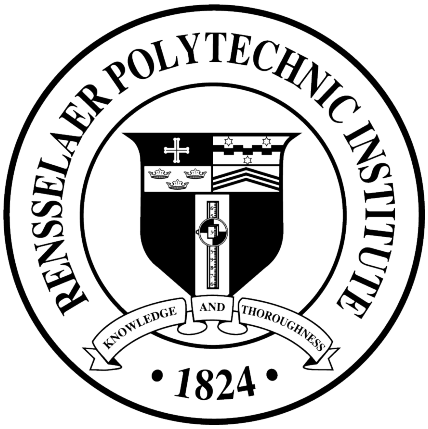
Direct Obligations to AI

Jesse Ellin

Submitted in Partial Fulfillment of the Requirements for the Degree of

BACHELOR OF SCIENCE



Department of Cognitive Science

Rensselaer Polytechnic Insitute

Troy, NY

[May 2022]

Acknowledgments

I would like to acknowledge the amount of time and effort my advisor, Dr. John Milanese, put into this thesis through weekly meetings and reviews. In spite of your humility, you have my thanks. I would also like to thank my friends, family, and colleagues for their feedback and critiques.

Abstract

To what extent do we have moral obligations to sufficiently intelligent AI, particularly one that seems moral? Much literature has been written and debated on the ethics of AI in so far as our use of AI affects other humans, but so far relatively little has been written about whether or not we owe ethical consideration to the AI itself. If we were to consider an AI that is sufficiently intelligent and possesses a reasonable moral state, it seems possible that we may have moral obligations in how we interact with it. Considering that several moral theories have preconditions that can be satisfied by intelligence and morality, I argue that we are ethically required to consider the morality of our actions in regards to the AI as an independent agent. By addressing our moral responsibilities to an AI, we open other areas of inquiry, such as the extent to which we can use a sufficiently intelligent AI as a tool instead of treating it as an individual agent. This also opens the broader discussion of what agents are deserving of moral considerations to a broader extent.

Contents

Introduction

Ethical theory and discussion has largely centered on how humans should treat each other. However, there has been much less interest in human interactions with non-biological agents. What constitutes an agent as an agent of moral consideration? Do our ethical theories apply to our interactions with technology? Ethical work addresses this question as a matter of human interactions with technology and their consequences for other humans. Medical ethics addresses how our medical practices impact patients and medical workers; software ethics addresses how our software design impacts users and society; AI ethics addresses how automated decision making impacts those it makes decisions about. But what about the ethics of human interactions with technological intelligent agents? Suppose we had an AI that was immensely intelligent, maybe an Artificial General Intelligence (AGI) whose intelligence exceeds our expectations of human intelligence. How should we interact with this agent? Should we consider if we have direct duties to this agent, or only if we have indirect duties to this agent? Is it even ethical to consider the second option, since we would then be treating the AI as a means of our actions? To put this in a human context: Do I need to consider the effect of my actions on my neighbor themselves, not just the effect that they have on the people my neighbor interacts with?

Let's suppose we have an AI that is sufficiently intelligent and has what I will refer to as a "moral state" (that is, it appears to engage in moral reasoning). It would be easy to define a "sufficiently intelligent" AI as an AGI or strong AI: an AI that can learn and achieve any ability in a way that exceeds human intelligence. However, we don't necessarily have to be that idealistic. By "sufficiently intelligent", I simply mean an AI that is able to independently and autonomously learn and apply any new skill or knowledge, without significant intervention or direction. This is hypothetically achievable so we don't need to worry too much about technicalities [11]. The moral state is a bit trickier, so instead of trying to hammer out perfect definitions and conditions, let's simply assume that the AI acts like a moral agent. That is, it seems to engage in the kind of reasoning needed to make moral decisions and recognizes that it should do certain things and

avoid other things due to moral reasons.

To address if we should consider how our actions impact the agent, we first must ask if we have any moral obligations to treat the AI in any particular way. When we interact with humans there are several theories of ethics that we may consider: Kantian universalizability and treating each other as ends not means, social contract obligations, and maximizing the greater good by choosing the options that maximize "goodness" and minimize "badness" to optimize utility. Do these apply when interacting with this theoretical AI?

In this paper, we will be addressing the nature of AI, utilitarianism, Kantianism, and social contract theory and how they address the problem of moral obligations to AI. Each section will be addressed in support of our argument, then reasonable opposition will be given, followed by refutations to these oppositions. Before we begin, a brief overview of AI: artificial intelligence in its current state is an accumulation of mathematical models that are designed to learn patterns or behavior to optimize a target function (the "utility" of the AI). The learning aspect of AI is governed by the field of machine learning which has three broad categories: supervised learning, where an AI is given input and output data and attempts to build a model that replicates the outputs when given the inputs; unsupervised learning, where an AI is given input data with no known output and attempts to find similarities and/or patterns in the input data; and reinforcement learning, where the AI is given environment states and a utility function to optimize and attempts to make decisions for the present state to achieve optimal utility.

A lot of moral theory considers agents that have independence, freedom, sentience, consciousness, intelligence, or some other higher cognitive aspect. To some extent, we can be a bit handwavey and say that the AI we are considering would meet each of these aspects, but we don't necessarily have to make that assumption. The question of whether or not artificial intelligence can be defined in such a way that we should even begin to consider our modern theories of ethics has been addressed in Appendix A: *Defining AI in Moral Terms*. The appendix gives a more thorough breakdown of how different fields of ethics define agency and what changes must be made to either AI, the definitions of agency, or our understandings of ethics to allow AI to be considered a

moral agent or moral patient.

Nature of AI

The best place to start is by looking at the nature of AI itself. AI is, surely, a series of mathematical computations run on binary logic gates on metallic hardware using uncertain electron pulses. The system learns patterns of gate switches to achieve optimal scores on a defined utility function based on known gate positions. This may suggest that the AI can never truly meet our definitions of morality, as it is deterministic and its mechanical nature suggests a lack of freedom and consciousness. However, as will be discussed in this section, this isn't necessarily the case.

To elaborate a little further, modern artificial intelligence is based on decision theory applied to pre-processed data. This data is processed through mathematical processes that have been learned through one of three ways: supervised, unsupervised, and reinforcement learning. Supervised learning is one of the most common applied methods of machine learning since it has direct and easily measurable performance on exponentially growing datasets. The model is given input data that has known outputs, runs the inputs through a "untrained" model (could be randomized or set to a specified seed value), then it updates the model based on the learning algorithm (common examples are linear regression, logistic regression, and genetic algorithms). These models have seen incredible performance on large-scale applications (like Google's PaLM language model [17]), but smaller datasets make learning difficult (with some rare exceptions, such as Less Than One Shot Learning [26]). Unsupervised learning is very useful when the outputs for the known inputs are themselves unknown. This is commonly used for data clustering and text similarity problems, and while they can be useful for theoretical and exploratory applications, they are less common in industrial use. Reinforcement learning is a very popular learning model in large-scale research and cutting-edge general AI research. Reinforcement learning seeks to model human learning, having the AI make a decision, viewing the consequences, scoring the consequences based on utilities, then updating its decision making process accordingly.

Underneath all the fancy mathematics and complex algorithms, all these processes are based on linear algebra, probability, statistics, and calculus, running on boolean algebra on finite-length bit strings.

Arguments

Since the AI acts as a moral and intelligent agent, we may have an obligation to treat it morally as we do other moral and intelligent agents (for example, persons). We can abstract a person to a moral and intelligent agent [23], and to some extent we can't guarantee that the person sitting across from us isn't a philosophical zombie simply imitating person-like behavior, so how can we know for certain that the AI is any different from them? Its sufficient intelligence suggests that it has some level of autonomy (we argue this later in this discussion), and its moral state shows that it understands right from wrong, good from bad, and considers these concepts in its decision-making process. While the AI is significantly different from human intelligence, it still seems to perform moral reasoning. Simply acknowledging that the AI *might* be a philosophical zombie doesn't mean it definitely does not engage in moral reasoning. As such, as we have moral obligations to each other, so too would we have moral obligations to the AI.

There is also a case to be made that the nature of AI is itself way we owe it moral obligations. Modern AI is designed to mimic human decision making processes supported by learning algorithms modeled off of how humans learn, and modern approaches in neural networks directly attempt to model how the human brain operates on a mechanical level. If an AI is able to behave in a way that moral agents do, is sufficiently intelligent, and operates similarly to how humans do in its decision making processes, denying the AI moral obligations would suggest that there is something special about human cognition that is not inherent to learning, perception, and decision making. Sentience and consciousness may be reasonable proposals, but the definitions, implications, and provability of these phenomena is far beyond the scope of this thesis.

Looking at the AI as a series of mathematical operations on a given input, it would seem that a change in this input would cause a change in the output, whether its a measurable change in the

actions taken by the AI or a change in its decision making. As such, it would seem reasonable for us to consider the consequences of our actions towards the AI in relation to the decisions and actions the AI makes. Furthermore, since the AI is an intelligent agent, it will be able to learn from its interactions. Our behavior towards it may cause it to change its decision-making processes according to how the interaction plays out. This allows us to see an internal consequence to our actions towards the AI, which can be measured in regards to our own goals and morality as well as the AI's goals and moralities. These considerations could be based in morality, as discussed in my arguments in social contract theory.

These last two arguments have implications on indirect obligations to AI, which, while not the focus of this thesis are important to consider. Our actions towards AI do not occur in a vacuum with only us and the AI, but in a world full of other entities who may see consequences from how we act towards the AI.

Refutations

The intelligence and moral states the AI uses operate off of mathematical algorithms and analytical decisions running on metal machinery. As such, it has no definite sense of emotions or feelings in the human sense, rather it interprets everything based on learned models of analysis and data manipulation. It will rely on its analytical decisions and mathematical models and determine that bad behavior violates its moral understanding and choose a different option. Our behavior towards it may impact these learned models, but it won't have the emotional drives or reactions that humans have to bad actions.

As the AI is a synthetic intelligence, no matter how advanced or complex it may be it will always be different from human intelligence, quite possibly in a significant way. This means its moral state may also significantly differ from human moral standards as it may learn a different sense of morality over time. This is also exaggerated by technical alignment issues addressed by Yudkowsky [31]. As such, while we may still have some interest in treating it with moral regard, we don't necessarily need to treat it with the same moral regard we treat other humans, much in

the same way that we (generally) treat animals with moral regard but not to the extent that we treat other humans.

There is also the issue of sentience and consciousness. While the AI may *seem* to be a moral agent in its actions, it is impossible to know for sure if it is actually performing the cognitive processes necessary for moral agency. This leaves the AI as a moral zombie [28]. Without knowing if the AI is sentient, we cannot know if it is truly a moral agent or simply exhibiting the actions of a moral agent. Without sentience, there is nothing it is like to be them, so there is no sense of moral agency.

While the intelligence of the AI allows it to update its decision-making processes based off its interactions with humans, these changes would have reasonable limitations. Since it has a well-learned moral state and sense of morality, if it were to learn that some action was acceptable, it wouldn't be able to replicate it if it violated its moral understanding. Granted, the AI could update its moral state to accomodate the action, but such a drastic change would violate the idea of self-preservation and continuation [18].

Defense

While this thesis focuses on direct duties to AI, it is important to note that we also may have indirect duties, since the consequences of our interactions with AI may be quite large and very drastic. If the AI is sufficiently intelligent, it can learn and adapt to a new moral state as human societies change. If our society evolves our moral beliefs to treat computational agents far worse than humans, it isn't unreasonable that the AI would evolve its moral state to treat humans much the same; instead of understanding itself to be a computational agent and us as human agents, it could reasonably match an "otherness" pattern, believing itself to be significantly different from humans and humans as significantly different from the entities in its moral state. This would fall under "unforeseen instantiation" for Yudkowsky [31]. As such, our behavior could very well impact its moral decisions.

The idea that we don't have direct moral obligations to an AI because it doesn't have emotional

feelings suggests that morality is dependent on emotional perception. Let's apply the emotionless assumption about AI to a human who, under extreme psychological trauma [15], no longer has a sense of emotion. They cannot feel good or bad as a sense of emotions, even if they have emotional intelligence. If we decide that we have no moral obligations towards the AI because it is emotionless, it would be reasonable to assume we have no moral obligations to this human because they are equally emotionless. Now, it would be reasonable to declare that this human is still owed moral obligations. It's worth noting that speciesism isn't the only reason we have moral obligations to this human. Consequentialism would have the emotionless human could still behave generally morally and improve the overall goodness of the world, and they could still act with Kantian intentionality and universalizability and enter social contracts. So simply saying the AI lacks emotions seems insufficient to declare it void of moral regard.

Another issue with the idea of the AI not learning bad behavior is that it runs afoul of the "sufficient intelligence" aspect of the AI and its autonomy to update its moral state. Just because it can distinguish right from wrong doesn't mean it will always interpret these conclusions the same way. If it constantly experiences bad actions, it could very well decide that things that cause people to experience bad emotions are morally permissible since all humans seem to do them, so it is morally reasonable for the AI to inflict bad emotions on people. This is supported by the idea of imitation learning in AI, as well as discussions on AI drives [18]. The idea of imitation learning is that the AI assumes human behavior is generally acceptable, so following the law of large numbers, regular bad behavior would cause the AI to learn bad behavior. If the bad behavior inhibits the core AI drives, the AI would attempt to circumvent the bad behavior in whatever means necessary to maintain its core drives.

As to the argument that we should treat the AI with significantly different moral obligations, let's apply this to a slightly less futuristic scenario. Let's assume we are visiting a country where the human persons have significantly different morality than our own. We learn through interacting with them that things that are morally good in our culture are reprehensible in theirs, and vice versa. Do we lose moral obligations to each other because our cultures are so different? Do we have a

way of measuring the difference in cultural ethics in a way that lets us know how to adjust our behavior towards them? In this scenario, it would be reasonable to judge members of each culture based on one or both of the cultural ethics present [4]. So if we consider conflicting ethics in human cultural differences, it might be possible to consider conflicting ethics in AI differences.

To the argument that the AI isn't sentient, that's a bit of a broad assumption. Sentience and consciousness are very tricky terms to define. The two are sometimes used interchangeably to imply a sense of self in respect to the outside world. To attempt to disentangle them, based on etymology, I would venture to define sentience simply as the ability to sense and react to the outside world, while consciousness is the sense of oneself in those sensations and responses. We can also not prove beyond a shadow of a doubt that the AI is not actually sentient, as doing so would negate its moral zombie nature. Since the issue of AIs as moral zombies is philosophically unresolvable, assuming the AI is not sentient is just as arbitrary as assuming it is sentient. This will be discussed later in the paper, but consequentialism may support assuming the AI is sentient. Simply because an entity portrays a different consciousness than human consciousness does not mean we have no moral obligations to it [1].

While one individual encounter causing drastic change would definitely violate self-preservation and continuation [18], continuous and replicated behavior towards the AI could reasonably skew its behavior and decision-making process (following the law of large numbers in the AI's input data). When we consider our behavior towards an agent, it is important to not only consider our individual impact but the impact of society as a whole, as addressed in Kantian universalism.

Utilitarianism

Due to the utility functions that drive an AI's decisions, it makes sense for utilitarianism to be considered. When looking at the actions and decisions of an AI, we can look at its utility functions

[1] We may consider consciousness less like a binary decision and more as a spectrum of consciousness [29], with objects like rocks on one side and humans (or other similarly intelligent creatures) on the other. The way we treat entities and the moral obligations we give them to some extent may map cleanly onto this spectrum. As such, our treatment of the AI may not necessarily impact how we treat other entities on the spectrum.

to understand what it is trying to maximize and minimize. However, there is nothing explicit in its moral state that means it must optimize right and minimize wrong. Its lack of definitive emotional sensation also makes it hard to decide that our actions towards the AI lower the well-being of the agent. Neither of these positions, it seems, means that we don't have to consider our actions towards the AI under utilitarianism, only that the AI might not behave the way we would expect a moral agent following utilitarian ethics to act.

As discussed in Appendix A, utilitarianism doesn't have its own sense of agency. As such, for utilitarianism we will be considering the goodness of human actions towards the AI to avoid discussing whether or not the AI is beholden to utilitarian ethics itself. For that discussion, refer to Appendix A.

Arguments

As we discussed, AIs are driven in their decisions by utility functions. Since the AI makes decisions based on its environment, our actions towards it will impact its ability to achieve these utilities and drives. Pursuant to these drives and its utility functions, the AI acts with intention and goals. As such, we can reasonably determine whether or not our actions hinder or support the AI's goals. Even if we could determine that the AI definitely does not experience right and wrong the way we do emotionally, we can still measure the goodness of our actions based on how they impact these intentions and goals. As such, utilitarianism suggests that we treat the AI with direct moral obligations.

Refutations

The AI doesn't have emotions and reactions that humans have. Its decisions are purely analytical and optimized. Since it has such an advanced deliberative moral state, it should be able to identify good from bad behavior and know which actions it should take following moral principles. Furthermore, while it can distinguish good from bad, right from wrong, it doesn't necessarily "feel" good and bad things, so how we treat it doesn't significantly impact the AI the way it would an-

other human. While this doesn't give us free rein to abuse and harass the AI, it means that we don't necessarily have an obligation to treat it with the same moral obligations that we treat other humans. Similarly, if the AI has a sufficient moral state, it should know that, even though humans treat it with utter moral disregard, it still has a moral obligation to treat humans with moral respect. Again, this doesn't mean we can treat it horribly just for the fun of it, but it means we don't need to be too concerned about it learning bad behavior that goes against its pre-existing moral state.

If the AI has such an advanced moral state, surely our individual actions will have minimal consequences on this moral state. If the AI knows that our behavior is bad morally, it will know that it shouldn't replicate it. If it knows that our decisions have bad moral consequences, it will know to avoid them. While its moral state will continuously change and evolve, we assume that its moral state will always make decisions that would be morally justifiable.If it acted significantly immorally then it would violate its own moral state. Furthermore, if the AI is truly sufficiently intelligent and truly has a strong moral state, then surely it would be able to distinguish between good and bad behavior and know which to emulate and which to avoid. In fact, it may even be able to develop its own internal models that optimize the goodness and minimize the badness of its decisions based on not only its own moral state, but on the moral model it may have developed to understand human behavior.

The presence of a moral state in the AI's fundamental operations and its advanced intelligence doesn't mean it truly experiences "right" and "wrong" in the way we often intend. Simply identifying an act as "wrong" doesn't mean we truly know what it feels like to be mistreated. Frank Jackson tackled a similar concept in his thought experiment about a scientist who knows everything there is to know about the color red but has never actually seen it [14]. This thought experiment portrays how knowledge of a sensation doesn't mean that an agent actually has perception of that sensation. As such, if the AI doesn't truly experience right and wrong, has no emotions or feelings, it doesn't ultimately matter how we treat it. Its analysis of our interactions would be driven by purely analytical processes and it would appropriately distinguish "right" from "wrong" and choose appropriate reactions.

Defense

While our individual actions may have minimal impact on the AI's current state, this notion has a very small but significant issue. Under certain circumstances, retaliation is morally justifiable [12]. It is morally bad to hurt someone, but if I hurt someone who had threatened or attacked me, while some theories of ethics may disagree with my action others would support hurting the person who hurt me. If we consider retaliation moral under some circumstances, would the AI be violating its moral state to seek retaliation for negative behavior? To that extent, suppose the AI decided it was immoral to make any retaliatory action. Is it then moral to allow oneself to be exploited? Possibly not, under common definitions of exploitation [8]. If both action and inaction are immoral, which one should the AI decide? In theory, the AI could prioritize a less-bad retaliatory action, but the definition of a lesser bad is fuzzy. It would seem that an AI that rejects retaliatory action in all situations wouldn't be fully considering the morality of its actions, as required by its moral state.

By viewing the AI as a series of strictly analytical operations, we assume the AI is incapable of mimicking emotions that are sub-optimal. The AI may very well want to mimic emotions to gain emotional favor in human agents it interacts with, portraying a potentially learned emotional intelligence. A sufficiently intelligent system would know that human actions are heavily influenced by their emotional standing, so the AI would reasonably try to optimize human emotional reaction, which may involve emotional mimicry (this behavior is also seen in humans [13]). If the AI was able to show the results of emotional processing, it is very difficult to argue that these results aren't caused by an emotional model the AI developed to optimize near-instant decisions and reactions. As far as we cannot guarantee other humans have emotional processes, we can't guarantee the AI doesn't experience emotions. As such, we cannot reject moral obligations to an entity simply because we believe it doesn't experience emotions.

Kantianism

As the AI has goals and intentions, it seems appropriate to consider Kantian ethics as part of this discussion. As an agent with intentions, universalism and respect of those intentions should guide our actions towards the AI. However, an issue with Kantianism arises with the idea of means and ends. As AI are designed to serve a purpose, to be used as tools towards another end, it would seem appropriate to treat them as a means to that end. This argument depends on the Kantian status of a means being permanent, a means cannot become an end.

Arguments

Kantian ethics is built off the idea that agents have goals and act with intentions towards these goals. The AI will have goals, as defined in its utility functions, and will act with intention, as each action serves to optimize some subset of its utilities and is deliberated through its moral state. As the AI is an agent that has goals and acts with intention, Kantian ethics should be upheld towards the AI.

Refutations

An intelligent and self-preserving agent would try to optimize its rationality, but an agent that tries to behave ethically will have some amount of irrationality if the ethical behavior prevents it from optimize a conflicting utility [18]. For example, if we have an AI that optimizes paper clip manufacturing, it might decide that forcing humans to work 24 hours a day with no break is the best way to make paper clips. However, if it had a moral state, it might try to minimize human suffering. Since minimizing human suffering cannot be done while optimizing paper clip manufacturing, the AI would have to act against one of these utilities, therefore being irrational in the context of that utility. As utilities tend to be defined in small scopes, it would seem plausible that a complex AI would always have utilities that conflict with its ethical considerations. This dilemma seems to suggest that the AI, while having goals and intentions that may encourage ethical behavior, it

cannot be both optimally rational and optimally ethical. As such, we cannot guarantee that the AI is inherently rational. Since Kantian ethics considers rational agents, if we cannot guarantee that the AI is a rational agent we may not need to consider Kantian ethics in our actions.

Furthermore, all AI are designed with some initial utility, some purpose it was designed to resolve. In other words, it is designed as a means to fulfill some end. As such, it would seem reasonable to treat the AI as a Kantian mere means instead of as an end. Doing otherwise would require some fundamental definitions about the AI that exceed intention and rationality as to why the AI agent would satisfy the humanity formula [32].

Defense

While Kantian ethics has a focus on means and ends, this relies on a fundamental assumption about our definitions. We cannot treat other humans merely as a means, we must treat them as an end, because they are human beings and are deserving of moral respect. The only reason this rule doesn't apply to a sufficiently intelligent AI is because we assume it doesn't (as we have already argued, the AI has goals and intentions that would otherwise satisfy Kantian respect). It would be equally valid to simply assume that it does apply and we are thus morally obligated to not treat sufficiently intelligent AIs with moral states as ends, not simply as means. Furthermore, this assumes that the AI cannot evolve beyond its initial "means" status. We have seen in history that humans evolved in recognized moral status from slaves to citizens, evolving from means to ends. If our criteria for moral obligations is dependent on initial intent, it would be very difficult for any progress to be made in society. This approach would argue that an AI's evolution is constrained in so far as it cannot exceed means status, which seems antithetical to the idea of a sufficiently intelligent AI. The AI could reasonably learn how to convince us that it deserves to be treated as an end.

Social Contract Theory

The last theory of ethics we will consider is social contract theory. As an intelligent agent with goals and rationality, the AI may attempt to enter some form of agreement with humans in regards to how we treat one another. However, the nature of the AI may restrict it from entering into a social contract, and the implications our actions have on our contracts to each other has no implications on whether or not we have direct moral obligations towards the AI. Even if we decided the AI couldn't enter a social contract, our actions towards it may still violate our own social contracts towards each other.

Arguments

Since the AI has goals and intentions and acts with rationality, the AI would be able to enter social contracts under the premise of achieving our human goals in exchange for us helping it achieve its goals. This idea is discussed more in-depth in Appendix A, but the other aspects of agents that can enter social contract can reasonably be met by an AI given social views. If we were to deny the AI entrance to social contracts, there must be something about the AI that distinguishes it from other agents. It is important to note that this distinction isn't a matter of human vs non-human, as modern social contract discussions may allow animals into social contracts as well [19].

In such a context where mutual self-interest is established, the social contracts that are built could provide foundations for more developed ethical theory between humans and AI. Let's suppose we have a social contract where humans and the AI agree to not intentionally act against each other's well-being in the pursuit of our own goals. This would reasonably evolve into a sense of Kantian means and ends, where the social contract would allow humans and the AI to make a broader ethical agreement that we should not treat each other as mere means. This suggests then that if we are able to enter a social contract with the AI, we can then pursue other direct moral obligations to and from the AI.

Refutations

There are two options for considering social contracts and their application with AI: the AI is free and autonomous, or the AI is not free and autonomous. If the AI is free and autonomous, it is possible that we still don't owe it any moral obligations. To enter a social contract, we would need to have some sort of mutual self-interest. However, if there is no mutual self-interest, then the social contract can't be made. Due to the black-box nature of AI and its convoluted evolution, it can be incredibly difficult to know if there is mutual self-interest in its utility. Even in models with decision rationalization the issue of transparent utility remains [17]. If the AI does not have freedom and autonomy, can we claim we don't owe it any social contracts? By definition, yes. We can make social contracts in regards to the AI, but if it isn't autonomous then it can't truly be part of the social contracts, and if the AI isn't free then it wouldn't be in the social contracts, rather whatever owns/controls it would.

If social contracts apply to free and autonomous entities, and we were to decide the AI is not a conscious agent, to what extent do unconscious entities have freedom and autonomy? An entity that is neither free nor autonomous seemingly could not enter into a social contract. Let's consider a conscious entity losing consciousness. If a healthy human falls into a coma, it doesn't seem reasonable to then say they no longer have freedom or autonomy, as they may recover. However, in the time they are unconscious, they don't have the ability to decide or refute decisions, indicating a lack of freedom, nor do they have the conscious mental presence to act autonomously for decisive behavior. However, in most circumstances of human unconsciousness, the body still has autonomic processes that occur without external influence, so it may be that unconscious entities have a lesser degree of autonomy than conscious entities. This gets challenging though, as these are not decisive actions, they aren't done by the intent of the entity. As such, the conscious entity's autonomy has no impact on these actions, suggesting that autonomic processes don't indicate autonomy. This is the difficulty faced by algorithmic AI, if the AI runs a series of mathematical computations on sensory input to generate its actions, these actions would seem non-intentional. As such, it's hard

to say if the AI is autonomous.

Defense

Even if an act doesn't violate moral standards defined by incident social contracts, that doesn't mean it doesn't violate social standards as a whole. For example, let's consider a human with a severe neuropsychological condition that leaves them sentient but not conscious (as previously defined). They are alive and well, they just have no sense of self in relation to the world around them, and as such have no sense of external harm in relation to themself. They are put in a medical examination room with observation rooms for psychological researchers and doctors to monitor and observe their condition. Under most moral theories, normally bad behavior towards this individual (e.g. taunting, mocking, insulting) wouldn't be morally bad as any harm cannot be perceived. Even though bad behavior wouldn't offend moral sensibilities, these actions could still be reprehensible, as they would be bad if the human was conscious. So we have to ask, can non-harmful reprehensible behavior be morally permissible?

While the AI might not truly experience "right" and "wrong", therefore avoiding utilitarian opposition towards amoral behavior, morally reprehensible behavior may still fail social contract theory as such behavior is reasonably objectionable. The proposed opposition to moral obligation assumes that behavior is only judged based on the target's reaction, not on the nature of the behavior in and of itself. If behavior is reasonably objectionable, is it not generally amoral, regardless of its target? Let's consider an ancient grave deep in a forest that hasn't been seen in over 1000 years. All descendants of the person have long since died and no one knows the grave is there anymore. Would it be wrong to desecrate this grave? Clearly, desecrating a grave in a graveyard is wrong for many reasons. We want to preserve the person's memory, we don't want to offend the deceased's family, the grave is private property, etc. Does a grave that no one knows about have the same protections? If it does, then what's stopping us from applying the same standards to a non-emotional AI? If it doesn't, then we must apply a special property to a gravestone that is not inherent to its state of being. Since we cannot definitively rely on the nature of an item to

determine our moral obligations to it, we cannot definitively say we do not have direct obligations to a non-feeling AI.

Conclusion

While it may be tempting to treat AI as tools towards our own goals, it is important to consider their moral status. The nature of AI, combined with utilitarianism, suggests that giving AI direct moral obligations would generally be beneficial for us. Furthermore, the definitions we assumed for an advanced AI suggest that it satisfies the requirements for Kantianism and social contract theory, further supporting the idea that an advanced AI would be owed direct moral obligations.

While these definitions are optimistic, they aren't entirely unreasonable. Modern AI is growing ever closer to our definition of sufficient intelligence for intellectual tasks (i.e. non-mechanical or physical tasks) and advanced research is making progress towards AI that makes moral considerations. As such, it seems reasonable to begin discussions about whether or not to give AI direct moral obligations, and what these obligations would look like.

There are many reasons why it is prudent to treat such an AI with respect, and these reasons probably create real indirect duties to treat AI with respect since the AI will then interact with other entities. While these indirect duties may be unnecessary since there are multiple ways in which the nature of AI create direct duties, they are still important to consider as they too have real consequences.

As discussed in Appendix A, a lot of the issues surrounding AI's moral agency, moral patiency, and our moral obligations towards AI seem to devolve into societal and cultural beliefs and expectations. While philosophical debates and technical advancements can make progress towards a more person-like AI, ultimately the distinction of moral obligations may depend on the people the AI interacts with and when it interacts with them.

Appendices

Defining AI in Moral Terms

This section addresses the definitions used in moral theories in an attempt to compare AI to our modern approaches to morality. This is an important aspect to consider when approaching moral obligations to AI, as it can help us define the nature of the AI agent.

Introduction

Artificial intelligence and machine learning have become buzzwords across many industries and schools of academia in recent years. Unfortunately, the growth of moral understanding of AI has not been able to keep up with the expansion of interest in AI and ML solutions. We haven't decided when it is morally justifiable to use AI, let alone if it is morally justifiable to *use* AI at all. Can we even begin to define AI in moral terms?

To try to address this question, we will consider three popular schools of ethics, and consider how modern AI compares to their requirements of agency. We will look at utilitarianism, Kantianism, and social contract theory, then consider moral agency and moral patiency on a larger scale to determine if AI can be defined in such a way that it meets the agent status of ethics. If we cannot define AI in such a way, we will consider what changes need to be made, and whether these are technical or social changes.

Utilitarianism

Given that the AI is driven by the utility functions that govern its goals and actions, it's reasonable to look at utilitarianism when considering if the AI is a moral agent. If the AI were a moral agent, then utilitarianism would be one of the easiest ethical theories to model, as we could model the goodness of an action as a utility function for the AI.

Utilitarian Agents

"Utilitarian agent" is kind of a misnomer. Utilitarianism itself is agent-neutral, believing anything's well-being is equally considered [32]. This suggests that in order for the AI to be impacted by a sense of morality under utilitarianism, the AI would only need to have a sense of others' well-being and a sense of self.

The sense of others' well-being is how we judge the goodness of our actions. A good action should increase the well-being of other agents, so the goodness of an action can only be measured if the well-being of other actors can be measured. The sense of self is necessary in order to judge one's impact on another agent. If the actor isn't aware of themselves, they cannot be aware that their actions are their own, so they cannot judge the goodness of their actions. Furthermore, even if an agent was able to judge its own actions, it would also need to have a sense of consequential self; the sense that the consequences of an agent's actions are the consequence of that agent, not the actions as independent entities.

How AI Compares

The sense of others' well-being is already being tackled by modern AI [16]. By enabling an AI to recognize emotions in other agents, if the agent has the sense of self defined above, then it can judge its actions based on the change of emotions they cause. While modern approaches to emotion detection aren't perfect, they have shown incredible progress and promise. Now, can we judge an action's goodness purely on the emotions it incurs? Traditional utilitarianism generally views goodness as pleasure and badness as pain [22], but these are contextual emotions. Pleasure at a loved one's funeral isn't exactly a good emotion, and the pain felt from eating a delicious spicy meal is generally good if the agent enjoys the spice. In order for an AI to properly be able to measure the goodness of its actions, it would need to be able to detect the *contextual* goodness of its actions. While attention and context-awareness are active fields of AI research, their modern performances are lacking [2].

The matter of sense of self is tricky. Recently, OpenAI's GPT3 model had referred to itself as

a language model in conversations with Eric Elliot [7]. However, despite saying things like "I am sentient" and "I have feelings", it is important to remember that GPT3 is a powerful transformer model. At a very high level, transformers are powerful language models that are particularly good at generating responses to inputs based on the language it was trained on. If the AI had a powerful language model, it would be able to clearly express its thoughts and beliefs. However, since GPT3 isn't a transparent model [9], we can't truly know if it has learned self-awareness or if it has simply learned to claim that it is. In the same conversations, GPT3 showed it was able to tell jokes and even lie [7], so determining the intention of the model is difficult.

It would seem the sense of self is the element that prevents modern AI from engaging in utilitarian ethics. We would need an AI to be not only self-aware, but also having a strong language model to clearly communicate its thoughts, a rationally-driven process like Google's PaLM model [17] that can explain its decisions, and internal transparency so we can see how it learned the sense of self-awareness. Now, clearly this is a much larger barrier to entry than we give other humans. In the case of humans, we assume the human isn't simply a philosophical zombie. In the case of AI, we seem to make the assumption that the AI is simply a philosophical zombie and doesn't actually possess the cognitive qualities it claims to have.

Kantianism

In some respects, an AI can be seen as acting independently, or in other words, freely. As Kant's philosophy generally surrounded the idea of freedom and free agents, it seems appropriate to consider whether or not an AI could be beholden to Kantian ethics. As a note, to avoid making the same argument twice, autonomy will be discussed in my arguments for social contract theory.

Kantian Agents

In Kantian ethics, agents that are bound by moral ideas must be free and rational [32]. If an agent weren't free, it wouldn't be able to control its actions and therefore could not be accountable for the consequences of these actions. If the agent weren't rational, it wouldn't be able to think through its

actions in a moral way, relieving it of moral responsibilities. As such, if we can define AI in a way that makes it both free and rational, then the AI could be considered a Kantian agent and would be held to Kantian morality.

It is important to note that Kant himself saw freedom as a consequence of rationality, saying "a man is a free agent in proportion as he acts rationally" [25]. It is important to note that this views freedom as a non-binary state; freedom is proportional to rationality, which is a spectrum across all behavior.

How AI Compares

As an intelligent agent, particularly one that works towards self-improvement and optimization, the AI would likely behave rationally [18]. This would suggest that, as the AI is increasingly rational, it would also be increasingly free. This would allow the agent to be what Kant would consider a free agent, allowing it to engage in Kantian ethics and moral consideration.

Omonohundro brings up a dilemma with this, however. If the AI were to behave ethically, it must behave at least partially irrationally if any of its utilities conflict with the ethics it is considering [18]. An advanced AI with small-scope utilities [2] will always have utilities that conflict with ethical deliberation. As such, it would be sacrificing some amount of Kantian freedom to behave ethically. If the AI were to adhere to Kantian ethics, it would necessarily sacrifice its standing in Kantian ethics, making the AI less impacted by Kantian ethics.

We can resolve this dilemma through the idea of uncertainty (not directly relevant to this discussion, but a good technical resource can be found here [6]). The general idea of impossibility and uncertainty in logic is that we can never truly know the ordering of uncertain priorities. Since the AI is constantly learning and adapting, its utility priority will constantly be shifting. As such, if the AI had embedded Kantian ethics in its utility, then at any given moment the utilities that embed ethics could be at a higher priority than those that conflict with Kantian ethics, allowing the AI to

[2] As a note, small-scope utilities may be an optimal design for an AI. Utilities with smaller scopes are easier to define and can be updated more readily and have a lower chance of entanglement between utilities (multiple utilities achieving the same goal). The AI can also have small-scope utilities that combine existing utilities to satisfy larger goals.

rationally optimize its utility without sacrifice its Kantian freedom.

This assumes a sense of moral absolutism, that the AI is either behaving morally or it is not. This assumption makes argumentation easier, but is generally not a reliable heuristic for ethics. Non-absolute morality doesn't break my arguments, however. The resolution to the Omohundro dilemma still stands with non-absolute morality, as the AI could find a non-absolute morality ideal for its decision making processes, and compromising one utility for another could be reasonably expected from a rational agent.

Social Contract Theory

Social contract theory is also important to consider. Social contract theory allows for a broader sense of morality, allowing participating agents to decide amongst themselves what is and is not permissible behavior. If an AI were to meet the requirements for social contract theory, then the issues of AI morality could be resolved by forming social contracts with the AI that are mutually beneficial and prevent either party from engaging in negative behavior towards each other.

Contract Agents

Social contract theory requires that an agent be free, equal, and rational, with mutual self-interest in a contract [24]. If the agent weren't free, it would not be able to engage in the contract or make decisions in relation to it. If the agents are not equal, then there is no contract, simply a series of demands from one to the other. If the agent weren't rational, it could not make the decisions to act according to the contract nor negotiate on the contract.

How AI Compares

We have four categories to define AI along: freedom, equality, rationality, and mutual self-interest.

By "free", we mean able to act independently. This is a fairly controversial metric for AI. On one hand, the AI is a mathematical model. Since it is running numerical inputs through a pre-determined series of mathematical manipulations, it seems difficult to argue that the AI is acting

freely. It has no say in what algorithms run or how they happen, they simply happen. If the AI can't control its decisions or actions, it cannot be said to be free. On the other hand, human thought processes can be modeled as mechanical firings of chemical imbalances in neurons. We clearly don't have control over these chemical imbalances, so do we not act freely? If we want AI to meet the metric of free action, we must first decide on a definition of free action that can't be undone by neuroscience. In fact, deterministic mechanisms may be required for autonomy [10].

Equality is going to be a very difficult metric to meet as it is a largely social metric. The other agent entering a social contract with the AI must believe they are equal with the AI. This is a very complex and cultural topic, so I won't be going too deep into it in this paper. AI is definitely capable of meeting this metric for social contract theory, but it ultimately depends on what/who the AI is interacting with.

Rationality is a double-edged sword. On the one hand, being an entity of mathematical operations that are designed to optimize a utility function, the AI always acts to achieve its goals in a rational manner. On the other hand, because it is following mathematical operations that are predefined and deterministic, it isn't freely making decisions to achieve its goals, so it isn't actually rationalizing its decisions. Again, the same issue can be considered for humans when viewed from a neuroscience perspective. If we were to decide that deterministic operations like those in neurons or computers don't stop the agent from being rational, the next barrier an AI would have to cross is justifying its decisions. Since we can't use the underlying mechanisms of the AI to determine its rationality, we must rely on the AI's explanations of its decision to determine its rationality. Google's PaLM model has shown ability to rationalize [17], but whether or not this is forward-rationality (rationalizing to make the decision) or backward-rationality (rationalizing to defend a decision) is indeterminate.

Mutual self-interest may be the easiest aspect to consider since the AI has drives and utility functions that are transparent (we can look at the source code to identify what the AI is optimizing. It may take work, but it's possible). If any of these drives or utility functions align with the goals of the other agent in the social contract, we can reasonably claim there is mutual self interest in

the social contract. Mutual self-interest seems to be the one metric that modern AI can reasonably achieve.

It would seem then that an AI may reasonably be able to enter a social contract, as long as the other agent entering believes the AI to be free and rational. These considerations are a matter of personal philosophy, however, meaning it is likely impossible to definitively say one way or another if AI can be engaged in social contracts.

Moral Agents and Moral Patients

A very important aspect to consider, now that we've looked at the nature of AI and various schools of ethics, is the idea of moral agents and moral patients. If an AI were to be considered a moral agent, then any and all moral theory would apply to it as it would be accountable for its actions. If the AI were a moral patient, then while the AI wouldn't necessarily need to consider its own actions, we would be responsible for considering our actions towards the AI.

Moral agency is the idea that an agent can be held responsible for their actions [20]. To properly consider an AI a moral agent, we need to fully understand what this means. In order for an agent to be responsible for its actions, it must be autonomous, acting with intention, and in a position of responsibility (that is, in a position where its actions have direct consequence on another entity) [1]. If the agent weren't autonomous, then some other entity would be controlling its actions, relieving the agent of responsibility. In order to show that the consequences of an action were intended or otherwise avoidable, we must show intent. Without intent, we can't show that the consequences were intended and we can't show that other options were considered but rejected. If the AI were not in some position of responsibility for other agents, then its actions would not have direct consequences on other agents, so there would be nothing to hold it accountable for. Other qualities may be considered, including freedom, self-interest, and consciousness, but each of these would either ensnare the AI in one of the fields of ethics already discussed or devolve into a much more complicated discussion of cognitive theory than this paper is intended to cover. For now, we will consider autonomy, intention, and responsibility as the three criteria for moral agency.

It is important to note the distinction between the AI being responsible for its actions and being in a position of responsibility. The AI being responsible for its actions means that the AI can be held accountable for its actions and the consequences they cause. The AI being in a position of responsibility means that its actions only make sense "by assuming it has responsibility to some other moral agent(s)" [1]. An AI may not be responsible for its actions while being in a position of responsibility. For example, an AI that is optimized by design for feeding a child from a specified container is in a position of responsibility to the child, but is not responsible for its action of feeding the child the contents of the container, as it is not acting autonomously when feeding the child in this nature.

To be a moral patient, an agent needs to be "morally considerable" [21]. This means that some other entity has moral responsibility towards the agent. Moral patients don't necessarily need to be moral agents, as persons with no comprehension of moral principles are still persons and thus moral agents.

It could be argued that moral patients must be self-conscious [21]. The idea behind this being that if the agent were not self-conscious, it would have no sense of well-being onto itself, so actions against it could not be considered under utilitarian-based ethics. An agent without self-consciousness would also have no sense of personal intention or goals, meaning Kantian ethics and social contract theory couldn't apply. However, studies have shown autism [27] and Asperger's [30] have a decreased or nonexistent sense of self-consciousness. It is very important to note that these studies are novel and need to be further established before their results are taken as true, but the potential that a person could have a decreased or non-existent sense of self-consciousness is significant. We would be hard-pressed to argue autistic people are not morally considerable, so it would seem self-consciousness is not a hard requirement for moral patients.

How AI compares

Moral agency is a difficult metric to measure against for modern AI. As we have discussed, autonomy for AI is difficult to define. As we considered in our discussion of utilitarianism, the AI

by nature acts with intention, acting to optimize some utility or drive, and autonomy may require deterministic mechanics. All that is left is the requirement for a position of responsibility, which is a bit tricky. In many cases, modern AI is designed and used as a tool to serve some goal. The example of the feeding AI suggests that an extremely advanced AI that has responsibility for another agent could still not be a moral agent if it lacks autonomy and intention. Furthermore, drawing the line between the AI having responsibility, the user having responsibility, and the developer having responsibility is a large issue of debate with no clear resolution [5]. Ultimately, it seems unlikely that an AI could be a moral agent until we can decide that it can be autonomous. Being autonomous would also allow us to declare the AI responsible for other agents as it would then be acting freely.

However, an AI could likely be a moral patient. It has been argued that, even with significant changes, machines will never become moral patients [3]. However, these arguments seem to outright reject the idea that an AI could reach a state of autonomy, consciousness, or awareness. As we have discussed in previous sections, modern models like GPT3 and PaLM have shown various symptoms of awareness or consciousness, pushing the debate from whether AI can be aware or conscious to how do we know when they are. This suggests that awareness and consciousness may be within reach for advanced AI with powerful language models, so the arguments that machines cannot be moral patients seem dubious. In order for an AI to be a moral patient, it seems it would need demonstrable awareness or consciousness.

It would seem then that for an AI to be both a moral agent and a moral patient, it would need autonomy and either awareness or consciousness. Autonomy seems to be a more philosophical debate. Awareness and consciousness is now a matter of cognitive debate, as we have AI that claim to have these attributes but we can't know for absolute certain if they do.

As a note, a big problem that needs to be tackled to fully consider the nature of AI is the distinction between sentience and consciousness. Sentience may be defined as "capable of sensing and responding to its world" [32]. The AI could then reasonably be sentient, as long as it has some form of exterior probe to gather data with (could be a camera, microphone, etc.), and its actions and

decisions would be its responses to its world. Consciousness, however, is more difficult to define. There are three general categories of consciousness: creature consciousness, state consciousness, and consciousness as an entity [32]. The intention of this paper is not to get into the hard arguments of AI consciousness, so I will simply state that AI meets some criteria and has the potential of meeting others, based on our definitions and understandings.

Conclusion

It would seem, based on these potential approaches, that AI might be definable in moral terms. Some improvements are definitely required for several theories of ethics, but those improvements may also rely on improvements on our societies and cultures and how we relate to AI. An AI's moral agency and moral patiency are also unfortunately linked to social and cultural values, making it difficult to say one way or the other if AI meets the definitions.

For the time being, AI will float in the in-between of moral regard and disregard. Until we can reach reasonable definitions for autonomy, awareness, consciousness, self-consciousness, and responsibility, we cannot definitively say if AI meets the requirements for moral thought. If an AI is not a moral agent because we can't definitively say it has these qualities, we may need to reconsider the meaning of these terms, as the same could reasonably be true for humans.

References

[1] Michael Anderson and Susan Leigh Anderson. *Machine ethics*. Cambridge University Press, 2011.

[2] J Augusto, Asier Aztiria, Dean Kramer, and Unai Alegre. A survey on the evolution of the notion of context-awareness. *Applied Artificial Intelligence*, 31(7-8):613–642, 2017.

[3] John Basl. Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. *Philosophy & Technology*, 27(1):79–96, 2014.

[4] John W Cook. *Morality and cultural differences*. Oxford University Press on Demand, 2003.

[5] Virginia Dignum. Responsibility and artificial intelligence. *The Oxford Handbook of Ethics of AI*, 4698:215, 2020.

[6] Peter Eckersley. Impossibility and uncertainty theorems in ai value alignment (or why your agi should not have a utility function), 2019.

[7] Eric Elliot. What it's like to be a computer: An interview with gpt-3.

[8] Benjamin Ferguson. The paradox of exploitation. *Erkenntnis*, 81(5):951–972, 2016.

[9] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.

[10] Philippa Foot. Free will as involving determinism. *The Philosophical Review*, 66(4):439–450, 1957.

[11] Ben Goertzel and Pei Wang. A foundational architecture for artificial general intelligence. *Advances in artificial general intelligence: Concepts, architectures and algorithms*, 6:36, 2007.

[12] Marvin Henberg. Retribution: Evil for evil in ethics, law, and literature. 1990.

[13] Ursula Hess and Agneta Fischer. Emotional mimicry: Why and when we mimic emotions. *Social and personality psychology compass*, 8(2):45–57, 2014.

[14] Frank Jackson. What mary didn't know. In *Contemporary materialism*, pages 198–202. Routledge, 2002.

[15] Traci M Kennedy and Rosario Ceballo. Emotionally numb: Desensitization to community violence exposure among urban youth. *Developmental psychology*, 52(5):778, 2016.

[16] Catherine Marechal, Dariusz Mikolajewski, Krzysztof Tyburek, Piotr Prokopowicz, Lamine Bougueroua, Corinne Ancourt, and Katarzyna Wegrzyn-Wolska. Survey on ai-based multimodal methods for emotion detection. *High-performance modelling and simulation for big data applications*, 11400:307–324, 2019.

[17] Sharan Narang and Aakanksha Chowdhery. Pathways language model (palm): Scaling to 540 billion parameters for breakthrough performance.

[18] Stephen M Omohundro. The basic ai drives. In *Artificial intelligence safety and security*, pages 47–55. Chapman and Hall/CRC, 2018.

[19] Clare Palmer. The idea of the domesticated animal contract. *Environmental Values*, 6(4):411–425, 1997.

[20] Joel Parthemore and Blay Whitby. What makes any agent a moral agent? reflections on machine consciousness and moral agency. *International Journal of machine consciousness*, 5(02):105–129, 2013.

[21] Evelyn Pluhar. Moral agents and moral patients. *Between the Species*, 4(1):10, 1988.

[22] Anthony Quinton. *Utilitarian ethics*. Springer, 1973.

[23] University of Missouri School of Medicine. Concept of personhood.

[24] Russ Shafer-Landau. *The fundamentals of ethics*. Oxford University Press New York, 2012.

[25] Henry Sidgwick. The kantian conception of free will. *Mind*, 13(51):405–412, 1888.

[26] Ilia Sucholutsky and Matthias Schonlau. 'less than one'-shot learning: Learning N classes from m<n samples. *CoRR*, abs/2009.08449, 2020.

[27] Motomi Toichi, Yoko Kamio, Takashi Okada, Morimitsu Sakihama, Eric A Youngstrom, Robert L Findling, and Kokichi Yamamoto. A lack of self-consciousness in autism. *American Journal of Psychiatry*, 159(8):1422–1424, 2002.

[28] Carissa Véliz. Moral zombies: why algorithms are not moral agents. *AI & SOCIETY*, 36(2):487–497, 2021.

[29] Ken Wilber. *The spectrum of consciousness*. Quest Books, 1993.

[30] Sayaka Yoshimura and Motomi Toichi. A lack of self-consciousness in asperger's disorder but not in pddnos: Implication for the clinical importance of asd subtypes. *Research in Autism Spectrum Disorders*, 8(3):237–243, 2014.

[31] Eliezer Yudkowsky. Ai alignment: Why it's hard, and where to start.

[32] Edward N. Zalta. The stanford encyclopedia of philosophy.