The Moral Status of AI: What Do We Owe to Intelligent Machines? A Review

I. INTRODUCTION

Who—and perhaps *what*—exactly has moral status? Philosophers have attempted to explain and ground moral status for centuries, tracing back as early as Aristotle in 300 BCE (Arnhart 1998). Broadly, moral status can be defined as being morally considerable or having moral standing. An entity has moral status if and only if its interests morally matter to some degree for the entity's own sake (Jaworska and Tannenbaum 2018). The moral status of an entity helps us determine if and what kinds of considerations are owed to the entity. Questions about moral status are important in certain areas of practical ethics, such as animal rights, bioethics, medical ethics, and environmental ethics. For example, is it morally permissible to conduct experiments on mice? Do we do something wrong when we consume factory farmed animal products? Does an embryo have moral status, and how does the answer weigh on the permissibility of abortion? Until relatively recently, normative questions that hinge on moral status were limited to familiar beings—humans and non-human animals. Yet, increasingly, philosophers are expanding their considerations of moral status to try to account for novel and unfamiliar beings: artificial intelligence (AI).

Much of the discussion surrounding the ethics of creating and proliferating AI tends to focus on the consequences for humans, and wrongly omit considerations for AI, which are essentially the "research subjects" in AI research (Basl 2014). As more research is dedicated to creating increasingly advanced AI, it is important to consider what properties confer moral status, and whether AI, either now or some point in the future, possess these properties. Determining whether AI has moral status will have implications for how it is created and treated, and help us to understand what we "owe" to AI, so that we do not run the risk of committing a grave moral wrongdoing (e.g. torturing a sentient AI). This paper will review the literature on the moral status of AI, emphasizing that although a majority of philosophers agree that AI could plausibly have moral status based on capacities, there is disagreement about the specific degree to which such AI has moral status. I will begin by briefly defining and describing the specific kind of AI that is relevant to this review of moral status: artificial general intelligence (AGI).

Next, I will review the existing dominant approach to understanding moral status—Kant's sophisticated cognitive capacities view—in order to situate the current debate surrounding the moral status of AI. I will then describe and evaluate the two major themes that arose from the literature review: the capacity-based approach to grounding moral status and degrees of moral status (moral patient vs. moral agent). To conclude, I will offer that a capacity-based approach that recognizes AI as moral patients or agents depending on the context is the most appropriate account of the moral status of emerging AI.

II. ARTIFICIAL GENERAL INTELLIGENCE

There is widespread agreement that current AI systems do not have moral status (Bostrom 2011). As far as today's programs are concerned, we are justified in copying, terminating, deleting, or using computer programs as we please. Therefore, this review is concerned with a specific type of AI that does not yet exist: artificial general intelligence (AGI). AGI is an "AI system that equals or exceeds human intelligence in a wide variety of cognitive tasks" (Everitt, Lea, and Hutter 2018). Unlike today's AI, which can only solve narrow sets of tasks, AGI can think generally and achieve goals in a range of environments—similar to an adult human with normal cognitive capacities. But why bother theorizing about what we might owe to systems and machines that don't even exist? Some experts, such as tech inventor Ray Kurzweil, predict that we will create human-grade AGI as soon as 2029. Philosophers have conducted surveys of AI researchers' estimations for when AGI will be created; medians fall between 2040 and 2050 (Everitt, Lea, and Hutter 2018). Several organizations, including Google's DeepMind and Google Brain, are making rapid advances in "expanding the adaptability of algorithms to multiple problem domains" (Livingstone and Risse 2019). Institutions such as MIT and Stanford now offer courses on AGI and AI safety, suggesting that the prospect of AGI is increasing (Fridman 2019 and Sadigh 2019). Given these estimates and indicators that the creation of AGI is quite plausible in the near future, it is best to be prepared in understanding AGI's moral status before it exists, so that moral relativism recedes and a clear normative course of action emerges.

III. THE KANTIAN ACCOUNT: SOPHISTICATED COGNITIVE CAPACITIES

A discussion of the moral status of AGI, or any entity for that matter, is incomplete without an understanding of the most famous account of moral status: Immanuel Kant's sophisticated cognitive capacities account. For Kant, the moral world was divided into persons, beings who possess rationality and autonomy and therefore ought to be treated as means to an end, and things, beings and objects that lack those sophisticated cognitive capacities and can be used as mere means (Kant 1785). On this account, the only beings or persons with moral status would be human adults with full sophisticated cognitive capacities. Kant considered infants, the cognitively impaired, and non-human animals to be things, meaning their value is relative and that there is no duty to respect them as means to an end. Kant's account has rightfully received a wide range criticism for its exclusion of infants and the cognitively impaired (Jaworska and Tannenbaum 2014). Indeed, such exclusion seems to run counter to our intuitions that we owe more moral considerations to infants and the cognitively impaired. Prominent philosophers have also criticized the Kantian account for its exclusion of all non-human animals; certain nonhuman animals like Koko the gorilla have exhibited high levels of autonomy and rationality, yet are still considered to be things since their capacities fall short of those belonging to adult humans (Singer 2009).

Despite its issues with being overly exclusive, Kant's approach to grounding moral status continues to be one of the most common approaches, likely due to its intuitiveness: sophisticated cognitive capacities, which often include some sense of self, intuitively warrant moral consideration. The literature reviewed for this paper largely confirmed this; in order to ground the moral status of AGI, philosophers appeal to the Kantian account. If AGI possess sophisticated cognitive capacities similar to that of an adult human being, then AGI has moral status similar to that of an adult human being.

IV. A CAPACITY-BASED GROUNDING

As previously described, the majority of current literature agrees on the humancentric idea that if an AGI is very much like our own intelligence, then this AGI is our moral equal. Philosophers arrive at this conclusion using slightly different premises, and their applications also slightly diverge. Schwitzgebel and Garza (2015) offer the "No-Relevant-Difference

Argument" which goes as follows:

> *Premise 1*. If Entity A deserves some particular degree of moral consideration and Entity B does not deserve that same degree of moral consideration, there must be some relevant difference between the two entities that grounds this difference in moral status.
> *Premise 2*. There are possible AIs who do not differ in any such relevant respects from human beings.
> *Conclusion*. Therefore, there are possible AIs who deserve a degree of moral consideration similar to that of human beings.

Schwitzgebel and Garza offer an intentionally abstract argument; there is no commitment to what constitutes a "relevant" difference, so their approach is widely applicable. Such a relevant difference could assume a traditional Kantian view that focuses on cognitive capacities such as rationality, or a more utilitarian view that focuses on capacities to feel pleasure and pain. Their argument's conclusion is also intentionally weak, as the inclusion of "possible AIs" does not burden the argument with technological optimism or commitment to any specific type of AI (and for the purposes of this paper, would safely include AGI).

Similarly, Basl (2014) offers "The Easy Case: Human-Like Consciousness" which holds that if an AGI has a conscious similar to our own—has the capacity for pleasant and painful experiences, imagination, memory, critical thinking, aesthetic and emotional experience, and moral agency—then "on any reasonable normative theory, theory of welfare, and theory of moral considerability, this being will be our moral equal." Although Basl's account is more specific than the one offered by Schwitzgebel and Garza in that it specifies the relevant capacities and draws in theories beyond moral considerability, it operates on the same fundamental assumption: that AGI's capacities, if similar to our own, confer equal or very similar moral status. Bostrom (2011) takes the capacities-based approach further by defending it against two possible objections:

> *Principle of Substrate Non-Discrimination*

> If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status.
>
> *Principle of Ontogeny Non-Discrimination*
> If two beings have the same functionality and the same consciousness experience, and differ only in how they came into existence, then they have the same moral status.

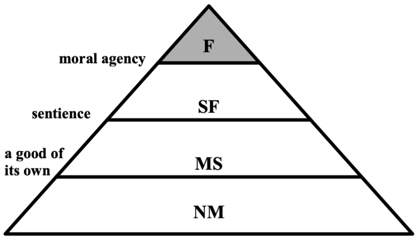
Once more, the author focuses on functionality (capacities) as a way to confer moral status to AGI, which certainly differ from us in substrate (silicon) and ontogeny (AGI is created artificially, not conceived and born in the way humans are). Supporting Bostrom's principles that defend a capacity-based approach against human-centric discrimination, Livingston and Risse (2019) offer functionalism as a way to confer moral status to AGI. Functionalism holds that the mind relates to the brain as software relates to hardware. Similar to the way software can run on different types of hardware, different types of physical entities can have minds (whether carbon or silicon-based). In this way, AGI is just a different entity with a very similar mind, and is therefore owed similar moral consideration as humans.

While differing slightly in premises and exact wording, the literature overwhelmingly supports that the most plausible theory of grounding moral status for AGI is capacity based.

V. DEGREES OF MORAL STATUS: PATIENTS VS. AGENTS

While there is agreement regarding what precisely grounds the moral status of AGI—capacities—there is disagreement about the specific degree of moral status. Philosophers understand moral status as a hierarchical, relational trait, rather than all-or-nothing (DeGrazia 2008). Scheessele (2018) offers a pyramid to understand the hierarchical degrees of moral status:

Fig. 1 The Moral Status Pyramid (MSP)



Source: Scheessele 2018

The pyramid is divided into four regions: (1) F for full moral status; (2) SF for significant-full moral status; (3) MS for minimal-significant moral status; (4) NM for negligible-minimal moral status. An entity that meets a particular threshold will also meet all of the lower thresholds. Philosophers mainly disagree over whether AGI would be in the "F" full moral status region or "SF" significant-full moral status region. An AGI with full moral status would be both a moral agent and a moral patient, meaning it is subject to moral obligations and other moral expectations in addition to having its interests considered for its own sake. Put simply, an AGI with full moral status would be held morally and individually responsible for its acts. Imagine that an AGI, while serving in the U.S. military, kills innocent civilians abroad; the AGI would be held responsible and punished only if we understand it to be a moral agent, not merely a moral patient (Sparrow 2007). The distinction between a moral patient and agent is thus significant.

Hakli and Mäkelä (2019) argue against the possibility of an AGI being held morally responsible for its actions. They contend that AGIs cannot be considered full moral agents because, as products of engineering, they cannot be autonomous in the sense of autonomy relevant for moral responsibility. Although AGIs and humans might be psychologically similar in all relevant respects, humans authentically acquire mental capacities and for that reason are autonomous, whereas AI acquire capacities by external manipulation and are for that reason not autonomous. It would be wrong to hold an AGI morally accountable for its actions in the same way we hold humans morally accountable for their actions, since AGIs lack the necessary autonomy for moral agency. Philosophers such as Sparrow (2007) also refute the possibility of an AGI bearing responsibility, simply because it would be too difficult to "punish" an AGI in any meaningful way. In order for an AGI to be capable of being punished, it must be possible for it to suffer in a morally compelling way; Sparrow argues that such a requirement is impossible. The conditions for moral responsibility and agency—autonomy and capacity for punishment—are thus too demanding for an AGI to be a moral agent.

Other philosophers such as Gordon (2018), Sullins (2006), and Laukyte (2017) argue that if AGIs are capable of moral reasoning and decision-making on a level that is comparable to humans, then we must consider AGIs not only as moral patients, but also as full moral agents prepared to bear the burden of responsibility for their actions. In response to the autonomy objection to granting AGI moral agency, both Gordon and Sullins argue that autonomy is not necessarily a precondition of moral agency, and that not all humans have autonomy yet are held morally responsible. Another objection to AGI being a moral agent is that they are simply "following a program," and thus should not be held responsible. Gordon responds that AGI can learn new rules—think generally—and therefore can ultimately be held accountable. Whether we consider AGI to be morally responsible and subject to punishment in the same way humans are will have significant impacts on our treatment of AGI and, equally important, AGI's treatment of us.

VI. CONCLUSION

This paper reviewed the literature on the moral status of AI, emphasizing that although a majority of philosophers agree that human-grade AI in the near future—AGI—will have moral status based on capacities, there is disagreement about the specific degree to which AGI has moral status. To conclude, I offer my own account of the moral status of AGI, drawing upon the work reviewed in the preceding sections. First, I assume the majority view that a capacity-based approach to conferring moral status to AGI is the most appropriate and robust account. If an AGI has similar cognitive capacities to our own, then we owe it similar moral considerations and must take its interests into account for its own sake; we cannot "kill," enslave, run experiments on, or treat an AGI without first evaluating it as if it has equal moral status to that of a human being. I recognize a potential drawback of accepting capacities as grounds for moral status; it could happen that at some point in the future, we confer a higher level of moral status to AGI than a cognitively impaired human. However, I contend that such an intuitively wrong conclusion could be addressed by taking a disjunctive approach to grounding moral status, by including other approaches such as person-rearing relationships that confer moral status (Jaworska and Tannenbaum 2014).

On the issue of what degree of moral status AGI will possess—whether we ought to treat AGI as a moral patient or moral agent—I assume a synthesis of the opposing views. There may be instances in which an AGI should be held accountable for its actions, depending on a host of factors, including who created it, why it was created in the first place, and the context of its "wrong" act. In other instances, it may be overwhelmingly clear that an AGI did not meet the threshold for autonomy or free will, and therefore would not be held accountable. As AGI becomes more advanced and our existing institutions work to adapt, we are best served by my proposed flexible approach that evaluates AGI on a case-by-case basis. Just as some humans can legally claim insanity or being a minor to mitigate a punishment, some avenues of explanation should exist for entities as novel as AGI. Such an abstract and less rigid approach reflects the rapidly changing, unknown nature of AI. As AI research continues, it will be increasingly important for more thought to be given to our treatment of AI, especially since AI may differ vastly from us in physical appearance, and history has shown that humans already struggle to treat other humans in an equal and just way. Thus, the discussion of what we owe to intelligent machines must expand outside of philosophy departments into research labs, government offices, and everyday minds.

References

Arnhart, Larry (1998). “The New Darwinian Naturalism in Political Theory.” Zygon: Journal of Religion & Science, 33(3), 369-393.

Basl, John (2014). “Machines as Moral Patients We Shouldn’t Care About (Yet): The Interests and Welfare of Current Machines.” Philosophy & Technology, 27(1), 79-96.

Basl, John (2014). “What to Do about Artificial Consciousness.” Ethics and Emerging Technologies. Palgrave Macmillan, London.

Bostrom, Nick and Eliezer Yudkowsky (2011). “The Ethics of Artificial Intelligence.” Draft for Cambridge Handbook of Artificial Intelligence.

DeGrazia, David (2008). “Moral status as a matter of degree?” Southern Journal of Philosophy, 46, 181–198.

Everitt, Tom, Gary Lea, and Marcus Hutter (2018). “AGI Safety Literature Review.” International Joint Conference on Artificial Intelligence (IJCAI).

Fridman, Lex (2019). “MIT 6.S099: Artificial Intelligence.” MIT. https://agi.mit.edu/.

Gordon, John-Stewart (2018). “What to we owe to intelligent robots?” AI & Society, 1-15.

Hakli, Raul and Pekka Mäkelä (2019). “Moral Responsibility of Robots and Hybrid Agents.” The Monist, 102(2), 259-275.

Jaworska, Agnieszka and Julie Tannenbaum (2014). “Person-Rearing Relationships as a Key to Higher Moral Status.” Ethics, 124(2), 242-271.

Jaworska, Agnieszka and Julie Tannenbaum (2018). "The Grounds of Moral Status." The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/spr2018/entries/grounds-moral-status/.

Kant, Immanuel, and Mary J. Gregor (1998). Groundwork of the Metaphysics of Morals. Cambridge, U.K: Cambridge University Press.

Laukyte, Migle (2017). “Artificial Agents among us: Should we recognize them as agents proper?” Ethics and Information Technology, 19(1), 1-17.

Livingston, Steven and Mathias Risse (2019). “The Future Impact of Artificial Intelligence of Human and Human Rights.” Ethics & International Affairs, 33(2), 141-158.

Sadigh, Dorsa (2018). “CS 521: Seminar on AI Safety.” Stanford. https://dorsa.fyi/cs521/.

Scheessele, Michael (2018). “A Framework for Grounding the Moral Status of Intelligent Machines.” Artificial Intelligence, Ethics, and Society, 215-256.

Schwitzgebel, Eric and Mara Garza (2015). “A Defense of the Rights of Artificial Intelligences.” Midwest Studies in Philosophy, 39, 98-119.

Singer, Peter (2009). “Speciesism and Moral Status.” Metaphilosophy, 40(3/4), 567-581.

Sparrow, Robert (2007). “Killer Robots.” Journal of Applied Philosophy, 24(1), 62-77.

Sullins, John P (2006). “When Is a Robot a Moral Agent?” International Review of Information Ethics, 6(12), 23-30.