Rights for Intelligent Robots?

A Philosophical Inquiry into Machine Moral Status

Kęstutis Mosakas

palgrave
macmillan

Rights for Intelligent Robots?

Kęstutis Mosakas

Rights for Intelligent Robots?

A Philosophical Inquiry into Machine Moral Status

palgrave
macmillan

Kęstutis Mosakas
Vytautas Magnus University
Kaunas, Lithuania

ISBN 978-3-031-64406-1 ISBN 978-3-031-64407-8 (eBook)
https://doi.org/10.1007/978-3-031-64407-8

This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Acknowledgments

As a monograph based on my PhD thesis “Moral and Legal Rights for Intelligent Robots,” defended at Vytautas Magnus University (Kaunas, Lithuania) on February 10, 2023, this work owes its greatest intellectual debt to the supervision and feedback of my former PhD supervisor, John-Stewart Gordon. His philosophical insights, attention to detail, and valuable suggestions continually drove me to delve deeper into the topic of robot moral status and rights—a process that significantly refined my abilities as a researcher and strongly shaped my understanding of ethics. Without John’s professionalism, support, and encouragement, this book would not have achieved its level of academic quality and may well not have materialized in the first place.

I would also like to extend my gratitude to Sven Nyholm for his feedback on the fifth chapter of this book, for providing me with opportunities to present my research at Utrecht University in the Netherlands, and for his support in preparing the proposal for this project. Likewise, I would like to thank Vladislav V. Fomin, David J. Gunkel, Julija Kiršienė, Edita Gruodytė, Aušrinė Pasvenskienė, Tomas Berkmanas, the “Integration Study of Future Law, Ethics and Intelligent Technologies” project team, and the philosophy department of Vytautas Magnus University for offering feedback on my research or supporting me in other ways, as well as the Vytautas Kavolis Interdisciplinary Research Institute (VMU) for providing me with employment for the final stages of this project’s preparation.

Furthermore, I am grateful to the editorial team of Palgrave Macmillan for expressing interest in this monograph and for their exceptional support throughout the publication process, and to Edit911 for their excellent language editing services. Last but not least, I thank my friends, family, and my wife Jessica in particular for their continued support.

This research was part of the “Integration study of future law, ethics and intelligent technologies” project (No. 09.3.3-LMT-K-712-01-0173), being executed at Vytautas Magnus University. The research was funded by the European Social Fund under Measure No. 09.3.3-LMT-K-712, “Improvement of the qualification of researchers through high-level R&D projects.”

Contents

CHAPTER 1

Introduction

Recent years have been marked by a rise of philosophical interest in questions related to artificial intelligence (AI), or “the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings” (Copeland 2020). AI systems are becoming ubiquitous. They are used for all kinds of tasks that humans are unable or unwilling to perform themselves, ranging from rather mundane ones, such as vacuuming, to something as impressive as space exploration. Indeed, it is anticipated that in the 2020s robots and AI applications will increasingly take on types of skilled or semi-skilled labor (Geospatial World 2020), and the possibility of full automation of work, according to numerous researchers, may well be within our reach in the coming century (see Grace et al. 2018). Although the transformative nature of AI comes with many challenges, it also offers an immense potential to ease human life and improve its quality. For this reason, we can be virtually certain that our future will be intertwined with AI.

Nevertheless, AI is still a new phenomenon in the history of humanity. Although the idea of machines becoming intelligent and capable enough to replace humans in various tasks has been around for many decades, the capabilities of AI systems exploded only in the 2010s. This was largely due to the highly potent machine learning model known as “deep learning,” which became feasible at that time due to advances in hardware. Currently, developments in the fields of AI and robotics are rapidly gaining momentum, and exactly how they will transpire remains an open question. As a

K. Mosakas, *Rights for Intelligent Robots?*,
https://doi.org/10.1007/978-3-031-64407-8_1

result, the future relationship between AI and humans is a nebulous and contested topic that requires further input from many different disciplines, including philosophy.

Some of the most fascinating prospects for upcoming technological developments involve robots—autonomous artificial agents with physical bodies that will work alongside humans and even share social spaces with them. Starting off as something like sophisticated tools in the early 1960s, robots have gradually evolved to emulate humans and animals, resembling them in various ways. Today, robots not only automate dull, dirty, and dangerous work but also entertain us and even perform certain social functions by interacting with us. This opens up the possibility of robots becoming something more than a raw labor force; perhaps they could become companions or even romantic partners (see Hauskeller 2014; Nyholm and Frank 2017)! The indeterminacy of our future with these artificial beings requires us to consider new philosophical questions as well as to revisit some of the old ones. How are robots different from humans and how should we think about human-robot relations? Could a machine experience emotions in the same way as humans do? Could it think? Could it become creative or broadly intelligent? Or is there a hard limit on the extent to which an artificial entity could instantiate these qualities?

In this work, I consider a related question: the ethics of the human treatment of robots and the possibility of robot rights. Although the central problems in AI ethics concern human welfare,[1] some scholars have begun to wonder whether machines could become morally considerable in their own right, in which case they would be entitled to our moral concern and protection. Understanding the moral constraints that should govern our treatment of robots is crucial, as their numbers are bound to grow considerably in the future, along with their involvement in various human activities. In the short term, this issue pertains to the impact that violent acts toward robots may have on human interactants. In the long term, the questions of robot moral status and rights may prove important as well, urging us to consider them thoughtfully in advance.

[1] The potential problems include automation of work and technological unemployment, concerns regarding machine bias and algorithmic opacity, issues surrounding privacy and surveillance, the allocation of responsibility in AI-based decision making, challenges related to human-robot interaction, ethical programming of autonomous AI systems, existential threats posed by the possibility of artificial superintelligence, and many others. All these issues have the potential to significantly undermine human welfare if left unaddressed, and that is the primary motivation for conducting research on them (see Coeckelbergh 2020).

While the idea of robots becoming so sophisticated as to be morally on a par with animals or even humans is fascinating, it may also seem unbelievable and reminiscent of science fiction. As Levy (2005, 393) remarked, "To many people the notion of robots having rights is unthinkable." Yet he also observed that the same was true for many other previously excluded entities that now enjoy moral and legal protection (ibid.). Therefore, it seems prudent to remain open-minded regarding the question of rights concerning future robots and to give it some thought—a position that is now embraced by numerous scholars, most notably Gunkel (2012, 2018a). Naturally, there are still skeptics whose arguments warrant consideration too. However, the domain of robot rights has garnered significant momentum in recent years. The floodgates have been opened wide and show no signs of closing anytime soon.

1.1 Research on Robot Rights and Moral Status: Criticisms and Justifications

Before we delve into the criticisms and justifications of the research question driving this book, it may help to present the topic in a broader academic context. The question of robot rights and moral status is generally recognized as part of AI ethics—a new branch of applied ethics focused on identifying and addressing the ethical issues raised by the use and development of AI systems. Some of the published material on robot rights appears in entries on AI ethics in the main online encyclopedias of philosophy—the *Internet Encyclopedia of Philosophy* (Gordon and Nyholm 2021) and the *Stanford Encyclopedia of Philosophy* (Müller 2020, albeit to a considerably lesser extent). Robot rights and the more general question of how robots ought to be treated are also sometimes discussed in the context of "machine ethics," a branch of AI ethics that 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).[2] Although the central focus of machine ethics is on endowing robots with the ability of moral reasoning, some scholars regard robot

[2] Machine ethics should be further distinguished from robot ethics (or "roboethics") and computer ethics. Robot ethics is concerned with the ethical aspects of designing and deploying robots, whereas computer ethics is the ethics of the human use of computers.

rights as a subbranch of machine ethics (e.g., Yampolskiy 2016, 140). I will treat them as distinct but related fields that, in my view, share certain important justifications.[3]

In the context of contemporary scholarship, the position occupied by research on robot moral status, rights, and their moral treatment is hard to describe unequivocally. Whereas some scholars are highly optimistic about the future of technology and see no problem with granting rights to artificial beings once they reach a certain level of sophistication, others continue to see technological artifacts as mere instruments at the disposal of humans. These contrasting attitudes are well conveyed by Gunkel (2018a) in *Robot Rights*, where he observes that while some view the rights question as contingent on certain qualifying criteria and whether robots could meet them, others still perceive the idea of robot rights as akin to Levy's portrayal in 2005—as unthinkable. To justify the research topic, therefore, I will briefly comment on the skeptical views expressed by some scholars.

One reason for skepticism regarding robot rights research is that it distracts us from more urgent ethical problems. Some argue that because robots are not expected to have moral status anytime soon, we should instead focus on more pressing concerns. For example, Floridi (2017, 4) claims that "it may be fun to speculate about such questions [robot rights], but it is also distracting and irresponsible, given the pressing issues we have at hand." Birhane and van Dijk (2020) have made an even stronger statement, arguing that robots are not the kind of beings that could be granted rights in the first place and that the robot rights debate focuses on what is more properly understood as a first-world problem.[4] Even more recently, Müller (2021, 585) has suggested that the question of machine

[3] Both aforementioned fields generally focus on ethical questions concerning hypothetical future robots that are significantly more advanced than current ones. In my view, one of the main contributions that research on robot moral status and rights can make to machine ethics is to clarify how intelligent robots should treat other machines, as well as the extent to which they should prioritize their own existence and structural integrity when confronted with ethical dilemmas.

[4] They write: "We argue not just to deny robots 'rights,' but to deny that robots, as artifacts emerging out of and mediating human being, are the kinds of things that could be granted rights in the first place. Once we see robots as mediators of human being, we can understand how the 'robot rights' debate is focused on first world problems, at the expense of urgent ethical concerns" (Birhane and van Dijk 2020). However, their position has received forceful criticism from Gellers (2020, 18), who regards it as "logically flawed and deeply contradictory."

moral agency and patiency "may not even present itself now" and that "it looks like a fiction for philosophers."

However, I believe there are at least three reasons why these skeptical attitudes are somewhat misleading and why we should take the question of robot moral status and rights seriously. The first and perhaps the most obvious reason is that research in this area will help us make better-informed decisions in the future as robots become more advanced. This is particularly important given that, as Glannon (2001, 3) rightly observes, our moral attitudes and laws consistently lag behind technological developments, creating a risk that unprecedented technological progress will occur without appropriate ethical guidance. Although artificial beings that we could call "intelligent robots" (IRs) are still at least a few decades away, the importance of the issue at hand is underscored by the fact that these beings with tremendous capacities will play significant roles in our society and routinely interact with human beings. Therefore, it is important to examine ahead of time how we can ensure that the way robots act and the way we act toward them are ethically appropriate and what moral and legal position robots should be accorded in our society more generally.

Of course, AI skeptics could insist that such guidance may never be needed. However, this view stands in tension with the generally optimistic attitudes of AI researchers and shuts down discussion of a potentially crucial issue in the future of humanity. I will substantiate this point further in Chap. 2, arguing that conducting research in machine ethics and related fields in advance is more prudent than ignoring these questions in the hope that we will never have to answer them.

The second reason for pursuing this research question is that it represents a useful thought experiment in moral philosophy. Even if we assume that robots will never acquire moral status—which is uncertain to begin with—engaging in discussions on the topic remains valuable. The questions of who and what matters morally and how we ought to treat other beings all pertain to complex and long-standing problems in moral philosophy. By reflecting on them in a different context—such as that of AI and robotics—we may uncover new insights and perspectives that can help to further elucidate these problems. So, while this book focuses on the moral status and rights of robots, the insights and arguments presented could be extended to or have important implications in other ethical contexts as well (e.g., animal ethics). Therefore, even if the optimism concerning the future of machine intelligence is overblown, research on the moral treatment of robots could still be a useful philosophical undertaking. This important point is commonly overlooked, possibly because the problem of machine moral status and rights is largely non-anthropocentric. That aspect distinguishes it from much of AI ethics, which is by and large concerned with the impact of AI systems on human beings. However, I believe that when we frame the problem in a broader philosophical context—i.e., in relation to fundamental questions concerning moral status, rights, duties, and values—its merits become more evident.

The third indicator of the value of researching this topic is the growing scholarly interest in it. According to a quantitative analysis of the academic literature on machine moral consideration (Harris and Anthis 2021), the growth in relevant publications has been exponential (see Fig. 1.1).

In the concluding section of their study, Harris and Anthis (ibid., 15) describe the current situation:

> Many scholars lament that the moral consideration of artificial entities is discussed infrequently and not viewed as a proper object of academic inquiry. This literature review suggests that these perceptions are no longer entirely

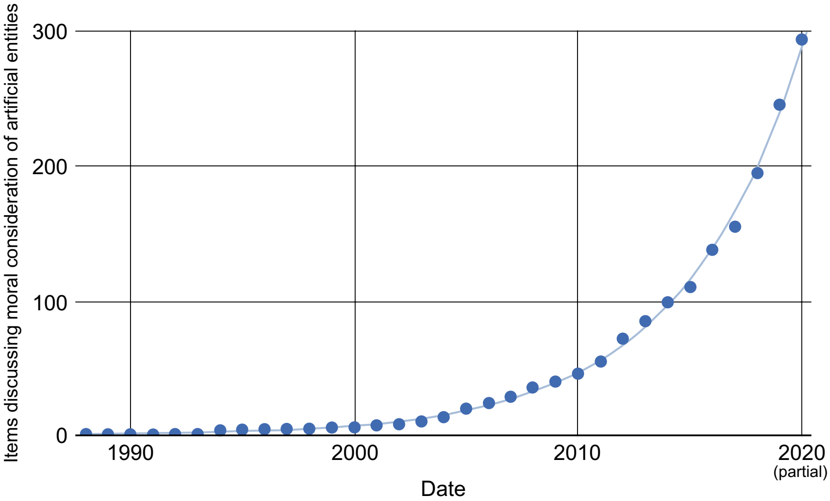

Fig. 1.1 The growth of academic literature on the moral consideration of artificial entities. Retrieved from Harris and Anthis (2021, 8), licensed under CC BY 4.0. No changes were made

> accurate. The number of publications is growing exponentially, and most scholars view artificial entities as potentially warranting moral consideration.

But if that is so, it seems more justified to perceive this field as new, promising, and one of growing importance, rather than fringe and irrelevant, as some skeptics may argue.

Overall, there are solid grounds for exploring the questions surrounding robot rights further. The possibility of IRs coming into existence is intriguing from a moral point of view. These machines are destined to transform the socio-political landscape and legal systems of human societies. While past technological developments have given us better tools to achieve our ends, it is not so clear that IRs will align with this pattern. Perhaps they should be seen as ends in themselves, rather than mere means to our ends! That is the timely and important question this work explores.

1.2 Contribution

With the rapid growth in the number of publications on robot rights in the last decade or so, diverse viewpoints have already emerged, and numerous related discussion points have been covered, at least to some extent. This includes the questions of machine moral agency (Sullins 2011; Véliz 2021), moral patiency (Tavani 2018), criteria for machine moral status and rights more generally (Sinnott-Armstrong and Conitzer 2021; DeGrazia 2022; Gordon 2021), the moral importance of artificial consciousness (Torrance 2014; Andreotta 2021; Ladak 2023), the epistemic challenge of detecting consciousness in robots (Agar 2020), indirect moral duties toward robots (Darling 2016), relational approaches to moral consideration in the context of robotics (Gunkel 2018b; Coeckelbergh 2010), and possible arguments against machine moral considerability (Torrance 2008). However, most contributions on these subjects have taken the form of academic journal articles or chapters in edited volumes. This book provides a more extensive treatment of the topic, akin to Gunkel's books

The Machine Question (2012) and *Robot Rights* (2018a).[5] It supplements the ongoing research with an in-depth analysis that touches on all the issues listed above, providing new insights while engaging with the latest academic literature.

Moreover, this work differs from Gunkel's books in several important respects. In his books, Gunkel surveys traditional accounts of moral status and rights and presents a critical examination of their underlying ways of thinking. According to Gunkel, the widely held view that one's moral status or rights depend on what one is (i.e., on one's ontological[6] properties) suffers from conceptual, epistemic, determination, and moral problems. Consequently, he suggests that it may be necessary to start "thinking otherwise" about ethical questions, which may help to avoid the perceived flaws in standard accounts. Although Gunkel does not claim to have a definitive solution, he appears to favor radical moral inclusion and the idea that moral consideration fundamentally hinges on relational dynamics rather than intrinsic properties.

In contrast, the present work reaches completely different conclusions from Gunkel's and pays attention to slightly different aspects of the problem. First, while Gunkel criticizes the standard properties-based view of moral status and rights, I affirm the moral importance of certain

[5] Although Gunkel's books focus almost entirely on machine moral status and rights, there are others that touch on these questions and are worth mentioning as well. *Rights for Robots* (Gellers 2020) discusses ethical aspects concerning the human treatment of robots, although its central focus is on legal rights and how animal and environmental law could inform our future decisions concerning robots. *The New Breed* (Darling 2021) draws on the history of human-animal relations and provides a few bases for treating robots well. However, it considers only currently existing and near-future machines that are nowhere near the level of sophistication necessary for moral status and rights. Nyholm's (2020) book *Humans and Robots* explores numerous issues concerning human-robot relations, dedicating one chapter to robot rights. Finally, *Killing Sophia* (Telving 2022) identifies some pitfalls in extending rights to non-conscious machines, although the discussion is simplified to enhance accessibility and the book is rather brief.

[6] In this work, the terms "ontological" and "metaphysical" will follow what I perceive as their standard use in contemporary philosophy, with metaphysics referring to the question of how the world is in the broadest sense and ontology addressing the question of what exists. Since ontology is a more specific branch of metaphysics, there is some overlap between the two, and therefore, in some cases, the terminology involved boils down to personal preference. For example, Gunkel (2018a) seems to prefer "ontological properties" when discussing properties like consciousness or rationality, whereas other scholars, like Danaher (2020), use "metaphysical properties" instead.

properties. For instance, I contend that phenomenal consciousness is necessary for moral status and that the properties of sentience and sapience (see Bostrom and Yudkowsky 2014) are (each) sufficient for it. Though this position is not new in ethics, I provide an extended defense for it by responding to new criticisms raised in the context of discussions on machine rights. I also raise several objections against the relational approach, which has been the main competing alternative to properties-based views.

Second, Gunkel presents our epistemic limitations regarding the presence of mental states in other entities as a stumbling block for properties-based views, leading him to explore more radical alternatives. By contrast, I see the epistemic problem as an inherent practical limitation of what is generally a highly plausible account. Therefore, rather than advocating for a new alternative, I dedicate considerable space to methodological considerations regarding our epistemic limitations in detecting mental states in other beings. I lend support to Agar's (2020) position that we can be justified in holding varying degrees of belief about whether different beings possess mental states. I maintain that both one's internal structure and outward behaviors must be considered when assessing the consciousness of a being.

Third, unlike Gunkel, who rejects the existence of morally relevant properties quite adamantly, I seek to integrate properties-based and relational criteria within a single framework of moral consideration. I argue that while some beings matter morally by virtue of possessing properties that are sufficient for moral status and rights, there are also entities with intrinsic moral worth based on relations (e.g., due to their symbolic value). This stance allows me to maintain that although robots may not acquire consciousness anytime soon, some of them could still attain a degree of intrinsic moral worth on relational grounds. This approach places my work somewhere between those who hold that robots cannot have intrinsic worth unless they become conscious and those who believe that certain current robots may already deserve considerable moral protection.

Finally, this work offers a more detailed approach to moral consideration. In his books, Gunkel deliberately attempts to turn standard moral reasoning on its head in hopes of finding a new, overarching ethical framework that would shake human exceptionalism to its core. While some elements of such a framework emerge in his analysis, the overall picture remains somewhat vague. In contrast, I outline different ways in which an entity could matter morally, introduce a terminology to differentiate them, and connect my analysis to practical considerations, such as possible moral grounds for extending legal protection to robots. Ultimately, while Gunkel and others may still critique the account defended here as "standard" and not sufficiently innovative, I attempt to remedy some of the shortcomings of the traditional viewpoint by broadening its scope, anticipating objections, and offering additional justifications. Thus, in broader terms, this book can be seen as a defense of traditional moral reasoning within the discourse on machine moral status and rights.

1.3 Methodology

Although I will cover certain methodological aspects more extensively in Chap. 3, it is important to note from the outset that the philosophical analysis in this work is pluralist in its normative assumptions. In other words, I will not approach ethical issues solely from the standpoint of one particular theory (e.g., utilitarianism, Kantianism, or virtue ethics). There is vast controversy in ethics as to which normative theory is right. As Muehlhauser and Helm (2012, 105) observe in the context of challenges related to building artificial moral agents, relatively simple ethical theories often yield counterintuitive implications, whereas more complex ones have been met with counterexamples of their own. They refer to this situation as "moral theory merry-go-round" (ibid.). Consequently, in recent decades it has become common to do moral philosophy in a way that is more flexible and sensitive to the controversy at hand. This includes doing justice to cases that resonate with our lived moral experience, even if they do not fit well within our preferred moral theory. It also entails not merely accepting outcomes when our preferred theory leads us to judgments that diverge from moral experience. The following point made by Warren (1997, 22) is of utmost importance here:

> To be credible, a moral theory must be reasonably consistent with "the common (and good) sense judgements that initially give rise to philosophical reflection on morals" (Hill 1992, 746). A theorist may be justified in rejecting some of the elements of common-sense morality; but in that case the theorist bears the burden of demonstrating that these elements are based upon errors of one sort or another—e.g. poor reasoning, false empirical beliefs, or ignorance of relevant facts. If none of the uni-criterial theories is sufficiently consistent with the elements of common-sense morality that we

> cannot reasonably be expected to jettison, then the goal of theoretical simplicity must be compromised for the sake of the equally important goal of adequately representing the moral data.

While it may be desirable to have one ethical theory applicable in all situations, approaching the problems in applied ethics in this "top-down" manner seems to underestimate the complexity of morality and creates the danger of "artificially forcing the issues to fit the theory," as Glannon (2001, 4) puts it. Ultimately, I agree with Warren that accounting for the existing moral data is no less important than theoretical simplicity. Hence, I will proceed from this basic assumption rather than starting from a specific moral theory in my analysis.

1.4 Outline

This book contains ten chapters, including an introduction and a conclusion. The next two chapters present important background issues and introduce the main concepts to be used throughout the book. More specifically, Chap. 2 provides additional detail on the importance of exploring AI-related ethical questions in advance, along with a working definition of "intelligent robot." Chapter 3 covers some of the main moral concepts involved, such as rights, moral status, and moral considerability. Additionally, it explains the basic aspects of rights, including their structure and function, and the key differences between moral and legal rights. It also highlights the importance of moral objectivism in addressing various problems that concern rights and morality in general.

Chapter 4 covers the notions of moral agency and moral patiency while also critically examining certain topics that have surfaced in the literature on AI and robotics regarding these concepts. It comments on issues such as "mindless morality" in the context of machine moral agency and on the animal-robot analogy in the context of machine moral patiency. Chapter 5 turns to some of the central problems analyzed in this work, including the necessary and sufficient conditions for moral status. More specifically, it provides support for the view that phenomenal consciousness is necessary for moral status, links the consciousness criterion with the properties of sentience and sapience (arguing that either of these properties is independently sufficient for moral status), and responds in depth to criticisms leveled against this properties-based view.

Chapter 6 transitions the discussion toward the relational dimensions of moral consideration. First, I outline the radically relational approach defended by Gunkel and Coeckelbergh and present five different objections to it. Then, I articulate a hybrid approach to moral consideration—one that considers both properties and relations as morally significant. I maintain that this approach better reflects moral data than "properties only" approaches and sidesteps the problems that the radically relational approach faces. Chapter 7 revisits issues related to consciousness. Specifically, it considers the epistemic problems involved, advocating for the use of probabilistic reasoning in discerning the presence of mentality in other entities. This approach is posited as more promising than its more restrictive alternatives.

Chapter 8 considers whether there are grounds for regarding robots as somehow morally lesser than biological beings, even if they were to possess the same properties that are necessary and sufficient for moral status. In this vein, arguments based on the observations that robots are non-organic, different in origin, weird, duplicable, non-human, or lacking special attributes such as a soul, free will, or dignity are examined and rejected. Chapter 9 reflects on possible moral justifications for providing robots with legal protection. It supports granting legal personhood to sapient machines while noting the risks related to moral underinclusion or overinclusion in cases of uncertainty. This chapter also considers the possibility of providing robots that possess human- or animal-like appearance or interactive capacities with some form of legal protection, outlining a few rationales for this step. Chapter 10 summarizes the main points of this work.

References

Agar, N. 2020. How to Treat Machines That Might Have Minds. *Philosophy and Technology* 33: 269–282.

Anderson, M., and S.L. Anderson. 2007. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine* 28 (4): 15–26.

Andreotta, A.J. 2021. The Hard Problem of AI Rights. *AI & Society* 36: 19–32.

Birhane, A., and J. van Dijk. 2020. Robot Rights? Let's Talk About Human Welfare Instead. In *AAAI/ACM Conference on AI, Ethics, and Society*, 1–7. https://arxiv.org/pdf/2001.05046.pdf

Bostrom, N., and E. Yudkowsky. 2014. The Ethics of Artificial Intelligence. In *The Cambridge Handbook of Artificial Intelligence*, ed. W. Ramsey and K. Frankish, 316–334. Cambridge: Cambridge University Press.

Coeckelbergh, M. 2010. Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics and Information Technology* 12 (3): 209–221.

———. 2020. *AI Ethics*. Cambridge, MA and London: MIT Press.

Copeland, B. J. 2020. Artificial Intelligence. *Britannica*. https://www.britannica.com/technology/artificial-intelligence

Danaher, J. 2020. Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Science and Engineering Ethics* 26: 2023–2049.

Darling, K. 2016. Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects. In *Robot Law*, ed. R. Calo, M.A. Froomkin, and I. Kerr, 213–231. Northampton, UK: Edward Elgar.

———. 2021. *The New Breed: What Our History with Animals Reveals about Our Future with Robots*. New York: Henry Holt.

DeGrazia, D. 2022. Robots with Moral Status? *Perspectives in Biology and Medicine* 65 (1): 73–88.

Floridi, L. 2017. Robots, Jobs, Taxes, and Responsibilities. *Philosophy & Technology* 30 (1): 1–4.

Gellers, J. 2020. *Rights for Robots. Artificial Intelligence, Animal and Environmental Law*. London: Routledge.

Geospatial World. 2020. Here Are the Top 5 AI Trends for 2020. Blog Post, November 6. https://www.geospatialworld.net/blogs/top-5-ai-trends-2020/

Glannon, W. 2001. *Genes and Future People: Philosophical Issues in Human Genetics*. Westview Press.

Gordon, J.-S. 2021. Artificial Moral and Legal Personhood. *AI & Society* 36: 457–471.

Gordon, J.-S., and S. Nyholm. 2021. The Ethics of Artificial Intelligence. *Internet Encyclopedia of Philosophy*. https://iep.utm.edu/ethics-of-artificial-intelligence/

Grace, K. et al. 2018. When Will AI Exceed Human Performance? Evidence from AI Experts. https://arxiv.org/abs/1705.08807

Gunkel, D.J. 2012. *The Machine Question*. In *Critical Perspectives on AI, Robots, and Ethics*. Cambridge, MA: MIT Press.

———. 2018a. *Robot Rights*. Cambridge, MA: MIT Press.

———. 2018b. The Other Question: Can and Should Robots Have Rights? *Ethics and Information Technology* 20 (2): 87–99.

Harris, J., and J. R. Anthis. 2021. The Moral Consideration of Artificial Entities: A Literature Review. *Science and Engineering Ethics* 27 (4, no. 53): 1–95. Creative Commons Attribution 4.0 International License. https://creativecommons.org/licenses/by/4.0/

Hauskeller, M. 2014. *Sex and the Posthuman Condition*. London: Palgrave Macmillan.

Hill, T. 1992. Kantian Pluralism. *Ethics* 102: 743–762.

Ladak, A. 2023. What Would Qualify an Artificial Intelligence for Moral Standing? *AI and Ethics*. https://doi.org/10.1007/s43681-023-00260-1

Levy, D. 2005. *Robots Unlimited: Life in a Virtual Age*. Boca Raton, FL: CRC Press.

Muehlhauser, L., and L. Helm. 2012. Intelligence Explosion and Machine Ethics. In *The Singularity Hypothesis: A Scientific and Philosophical Assessment*, ed. A. Eden, J. Søraker, J.H. Moor, and E. Steinhart, 101–125. Berlin: Springer.

Müller, V. C. 2020. Ethics of Artificial Intelligence and Robotics. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/ethics-ai/

Müller, V.C. 2021. Is It Time for Robot Rights? Moral Status in Artificial Entities. *In Ethics and Information Technology* 23 (4): 579–587.

Nyholm, S., and L. Frank. 2017. From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible? In *Robot Sex: Social and Ethical Implications*, ed. J. Danaher and N. McArthur, 219–243. Cambridge: MIT Press.

Sinnott-Armstrong, W., and V. Conitzer. 2021. How Much Moral Status Could Artificial Intelligence Ever Achieve? In *Rethinking Moral Status*, ed. S. Clarke, H. Zohny, and J. Savulescu, 269–289. Oxford: Oxford University Press.

Sullins, J.P. 2011. When Is a Robot a Moral Agent? In *Machine Ethics*, ed. M. Anderson and S.L. Anderson, 151–161. Cambridge: Cambridge University Press.

Tavani, H. 2018. Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights. *Information* 9 (4): 73.

Telving, T. 2022. *Killing Sophia: Consciousness, Empathy, and Reason in the Age of Intelligent Robots*. Odense: University Press of Southern Denmark.

Torrance, S. 2008. Ethics and Consciousness in Artificial Agents. *AI & Society* 22: 495–521.

———. 2014. Artificial Consciousness and Artificial Ethics: Between Realism and Social Relationism. *Philosophy & Technology* 27: 9–29.

Véliz, C. 2021. Moral Zombies: Why Algorithms Are Not Moral Agents. *AI & Society* 36: 487–497.

Warren, M.A. 1997. *Moral Status: Obligations to Persons and Other Living Things*. Oxford: Oxford University Press.

Yampolskiy, R. 2016. *Artificial Superintelligence: A Futuristic Approach*. London: CRC Press.

CHAPTER 2

The Emergence of Intelligent Robots and Machine Ethics

Will there ever be machines with human or even superhuman intelligence? If so, when should we expect them to emerge? Although the value of this work is not entirely predicated on whether such machines will in fact exist, these are nevertheless some of the key questions underpinning the study of machine ethics and robot rights, and therefore, they merit some discussion. Of course, no one can answer these questions with certainty, but some relevant perspectives deserve consideration, and that is the purpose of this chapter. Ultimately, I conclude that the likelihood of intelligent machines emerging at some future point is far from negligible and that therefore the ethical issues surrounding this possibility call for our attention now.[1]

2.1 What Is an Intelligent Robot?

Let us start by clarifying the concept of an “intelligent robot.” This may seem a rather straightforward task, but certain nuances concerning both “intelligent” and “robot” make it difficult to provide a precise definition.

[1] Some authors suggest that the expected impact of a superintelligent agent would be so high that it would justify careful exploration of machine ethics long before we have the capacity for developing one, even if the probability of such an agent ever coming into being is low (Shulman et al. 2009; see also Posner 2004; Rees 2003).

K. Mosakas, *Rights for Intelligent Robots?*,
https://doi.org/10.1007/978-3-031-64407-8_2

Therefore, my two main goals are: (1) to consider a working definition of an IR (albeit not exhaustive) that would suffice for the present discussion and (2) to explicate some of the nuances that make defining the concept somewhat difficult.

The first concepts to look at are those of machine and robot. Despite being used interchangeably in certain contexts, they are not in fact synonymous. According to Hall (2011, 29), the term "machine" can denote a device or apparatus consisting of motors, gears, and other parts functioning together to achieve some goal (e.g., a steam engine). But he also notes that this particular notion is too restrictive and fails to include other possible types of machines, such as abstract ones (e.g., the Turing Machine), virtual ones (a computer program), and other types (e.g., governments and legal systems)[2] (ibid., 29–30). What about the term "robot"? According to Bekey (2012, 18), a robot in the most basic sense is "a machine, situated in the world, that senses, thinks, and acts."[3] Thus, a robot is a particular kind of machine, so "machine" is the broader term whereas a robot is a machine with the properties indicated by Bekey. (In other words, being a machine is a necessary but not a sufficient condition for being a robot.) Notably, Bekey's notion precludes abstract and virtual machines from being regarded as robots, because neither of those are situated in the world, as he puts it. Additionally, Bekey admits that fully remote-controlled machines, as well as so-called telerobots,[4] are ruled out, because they are too dependent on human input and thereby do not "think." "Thinking" in this context presupposes a degree of autonomy: for a machine to qualify as a robot, it must be able to "process information from sensors and other sources, such as an internal set of rules, either

[2] Hall argues that although the government is composed of human beings, it has rules and laws by which these individuals must abide, and that "a person's whole function in the bureaucracy is to be a sensor or effector. Once the 'sensor-person' does his or her function in recognizing a situation in the 'if' part of a rule (what lawyers call 'the facts'), the system, not the person, decides what to do about it ('the law')" (Hall 2011, 30).

[3] "Thus," according to Bekey (2005), "a robot must have sensors, processing ability that emulates some aspects of cognition, and actuators. Sensors are needed to obtain information from the environment. Reactive behaviors … do not require any deep cognitive ability, but on-board intelligence is necessary if the robot is to perform significant tasks autonomously, and actuation is needed to enable the robot to exert forces upon the environment. Generally, these forces will result in motion of the entire robot or one of its elements" (cited in Bekey 2012, 18).

[4] Telerobots are machines that make only minimal autonomous decisions (Sullins 2011, 154).

programmed or learned, and to make some decisions autonomously" (ibid.). Various other issues related to what qualities a robot could or could not possess will arise in subsequent chapters—e.g., free will, intentionality, complexity, and consciousness—but Bekey's basic definition given above should suffice for our present purposes.

The other notion to be discussed, intelligence, has proven to be considerably problematic in several respects. First, the understanding of intelligence varies among different ethnic groups and cultures, as they tend to associate it with different competencies (Okagaki and Sternberg 1993; Niu and Brass 2011). To be considered intelligent in a certain ethnic group, one must possess skills that are valued by that particular group (Neisser et al. 1996, 80; Heath 1983). This variation in understanding intelligence has led some scholars to believe that it is a pure construct that is entirely contingent on one's social and cultural constraints with no objective basis (Berry 1974; Sarason and Doris 1979). Second, none of the attempts by different scholars to fit the varying conceptions of intelligence into a single, robust definition have received universal acceptance. For example, Legg and Hutter (2007) presented a survey of over 70 competing definitions of intelligence, featuring different accounts by psychologists, AI researchers, and collective definitions proposed by groups or organizations—all exhibiting considerable variation. Given this controversy, it would be unreasonable to advocate for a single "correct" definition here. Nevertheless, we need to devise a working definition for practical purposes.

Since the central theme of the present work concerns the moral status and rights of artificial entities, it may seem reasonable to only consider the definitions given by AI researchers. However, psychological definitions receive relative priority here, because "intelligence" has traditionally been considered in terms of how it manifests in biological organisms, such as human beings and, to a lesser extent, non-human animals. Nevertheless, in the context of robotics, not all these definitions are suitable. Some psychologists (e.g., Bingham 1937; Peterson in Sternberg 2000) incorporate terms like "organism" or "biological" in their definition, which seems to exclude the possibility of intelligent machines *a priori*. Certain other definitions contain references to culture. Gardner (1993), for example, describes intelligence as something that enables one "to solve problems, or to create products, that are valued within one or more cultural settings." With respect to "machine intelligence," then, the challenge is not only to accurately capture the common core within the varying conceptions of intelligence by psychologists but also to avoid biological, social, or cultural biases that some of these definitions exhibit. Furthermore, as Muehlhauser and Helm (2012, 103) highlight, "machine intelligence" should be distinguished not only from how distinct cultures and ethnic groups perceive intelligence but also from how laypeople typically understand it, since they commonly associate it with positive rather than negative traits.

When we turn to AI researchers, we still find no strong consensus on defining intelligence. Although their definitions do avoid the above-mentioned biases, certain differences remain. Perhaps the most noteworthy point of convergence among AI researchers is that intelligence seems to involve the ability to achieve certain goals or objectives (Albus 1991; Fogel 1995; Goertzel 2006; Horst 2002).[5] Some (e.g., Kurzweil 2000) have stipulated that an intelligent agent must be able to use its resources (including time) optimally in achieving its goals to qualify as intelligent, while others, such as Legg and Hutter (2006), add that intelligence concerns fulfilling goals in a wide range of environments. These aspects of intelligence are also captured by Muehlhauser and Helm (2012, 104), who refer to machine intelligence as "optimization power"—the greater the intelligence of a given machine, the more effective and efficient in using its resources it is in achieving its goals across a wide range of environments. (The notion of optimization power is invoked to avoid anthropomorphic biases related to the common colloquial use of "intelligence." Rather than adopting Bostrom's term "superintelligent machine,"[6] Muehlhauser and Helm (ibid., 104) opt to refer to a "machine superoptimizer" instead.) While according to Muehlhauser the notion of efficiency ("optimization power divided by resources used"—Muehlhauser and Salamon 2012; see also Yudkowsky 2008b) should play a part in defining intelligence, others have found this suggestion objectionable (see Legg 2009).

I think the definition proposed by Legg and Hutter (2006, 2007) captures the central aspects of machine intelligence while avoiding potentially unnecessary baggage associated with the concept. Legg and Hutter (2006,

[5] Some, however, choose to conceptualize intelligence in terms of adaptability in different environments through information processing (e.g., Wang 1995).

[6] "An intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills" (Bostrom 1998).

2) state that "intelligence measures an agent's ability to achieve goals[7] in a wide range of environments."[8] This is a viable candidate for a working definition. Of course, we cannot provide a brief yet accurate definition for such a complex concept without retaining a degree of vagueness. In this case, for instance, it is not entirely clear what qualifies as "a wide range of environments." Moreover, what about the goals? Would a machine still qualify as intelligent if its goals were so simple as to make achieving them in a variety of environments rather trivial? Perhaps Goertzel's (2006) definition of intelligence as the ability to achieve *complex* goals in complex environments might be better. But then the question of defining "complex" arises. These are all tough questions to answer, but the point here is that a comprehensive definition must consider a variety of aspects associated with machine intelligence: efficiency, complexity and broadness of goals, range of functional environments, adaptability, etc. For this reason, one may not want to stick to a single rigid definition but may instead prefer to take cognizance of the different aspects within the many existing definitions.

To further clarify the concept of an intelligent robot, we must say something about the well-known distinction between broad and narrow AI. Narrow AI systems, also commonly referred to as weak AI, only "demonstrate intelligence in one or another specialized area, such as chess-playing, medical diagnosis, automobile-driving, algebraic calculation or mathematical theorem-proving" (Pennachin and Goertzel 2007, 1). Of course, the single task that a particular narrow AI system is designed to solve may itself be extremely complex, but ultimately, such an AI system is

[7] Another interesting question is whether one can attribute a "goal" to a robot under the assumption that it has no phenomenal consciousness and therefore no intentional states. For the present purposes, I will set this discussion aside and simply stipulate that the notion of "goal" in this context can encompass something like apparent goals or "quasi-goals," as it were. For example, if a machine is programmed to play chess, winning the game becomes its presumed goal, regardless of whether it is conscious of winning or losing.

[8] Their proposed definition is formulated through a synthesis of numerous definitions by psychologists, AI researchers, groups, and organizations. Legg's and Hutter's analysis suggests that a general definition could be derived from three main features commonly shared by various definitions: (1) intelligence is a property that an individual agent has as it interacts with its environment or environments; (2) it is related to the agent's ability to succeed or profit with respect to some goal or objective; (3) it depends on how able the agent is to adapt to different objectives and environments (Legg and Hutter 2007). See also Legg and Hutter (2006) for a mathematical formalization of these common definitional features, aiming to develop a general measure of machine intelligence.

incapable of performing any other tasks. With regard to specific tasks, AI has already surpassed human beings in a number of challenging domains. Take complex board games as an example. In 1997, a chess-playing computer, Deep Blue by IBM, defeated Kasparov, a reigning world champion at the time, in a six-game chess match.[9] Another impressive demonstration of the capabilities of narrow AI systems is AlphaGo, a computer program that plays the Chinese board game Go, which is a more complex game than chess. In 2016, AlphaGo beat the world champion, Sedol, by using tree search to evaluate positions and selecting moves based on deep neural networks developed through learning from human experts as well as self-play (Silver et al. 2017a).[10]

Of course, the role of narrow AI is not limited to playing games. Today, narrow AI systems excel in numerous tasks, such as algebra, surveillance in advanced monitoring systems, detecting underwater mines, finding optimal routes in maps, traffic control, facial recognition, cancer detection, caring for the elderly and infirm, and many more (see Nilsson 2009). Additionally, the expansive use of deep neural networks, a highly potent form of machine learning, has made narrow AI even more efficient and expanded it to a multitude of new domains (see Heaton 2020).

Although narrow AI has been so successful in particular tasks, we are yet to see a machine with broad or general AI (also sometimes called strong AI)[11]—one that could "solve a variety of complex problems in a variety of complex domains, and that controls itself autonomously, with its own thoughts, worries, feelings, strengths, weaknesses and predispositions" (Pennachin and Goertzel 2007, 1). A machine described in this manner might seem remarkably humanlike, because it is proclaimed to possess traits typically associated with humans, such as being capable of

[9] Since 1997, chess-playing AIs have improved significantly and become virtually unbeatable by human opponents. In 2016, Kasparov admitted that "today you can buy a chess engine for your laptop that will beat Deep Blue quite easily" (Kasparov in Friedel 2016).

[10] Since 2016, several successors of AlphaGo have been developed, including AlphaGo Master, AlphaGo Zero, and AlphaZero, all of which surpassed the original version. Indeed, AlphaGo Zero was already beating AlphaGo after three days of training by 100 to 0 (Silver et al. 2017a), whereas AlphaZero has also mastered chess and shogi in addition to Go and has become a top player in all of them (Silver et al. 2017b).

[11] The term "strong AI" as originally used by Searle (1980) referred to a hypothetical AI system that could duplicate rather than merely simulate human cognition and understanding. Later on, the term was adopted by other scholars and used to refer to something different. For example, Kurzweil (2005) uses it to denote an AI system that exceeds the human level of intelligence.