KRITIKE VOLUME TWELVE NUMBER ONE (JUNE 2018) 182-200

Article

A Case for Machine Ethics in Modeling Human-Level Intelligent Agents

Robert James M. Boyles

Abstract: This paper focuses on the research field of machine ethics and how it relates to a technological singularity—a hypothesized, futuristic event where artificial machines will have greater-than-human-level intelligence. One problem related to the singularity centers on the issue of whether human values and norms would survive such an event. To somehow ensure this, a number of artificial intelligence researchers have opted to focus on the development of artificial moral agents, which refers to machines capable of moral reasoning, judgment, and decision-making. To date, different frameworks on how to arrive at these agents have been put forward. However, there seems to be no hard consensus as to which framework would likely yield a positive result. With the body of work that they have contributed in the study of moral agency, philosophers may contribute to the growing literature on artificial moral agency. While doing so, they could also think about how the said concept could affect other important philosophical concepts.

Keywords: machine ethics, artificial moral agents, technological singularity, philosophy of artificial intelligence

Introduction

Throughout history, technological advancements have led to philosophical inquiry. Developments in artificial intelligence (AI) research, for instance, have prompted philosophers to look into its surrounding foundational issues.[1] This paper examines one of the said issues, specifically focusing on the nature of artificial moral agency and how this relates to a technological singularity, the point in which AI will have greater-than-human-level abilities. Furthermore, this paper also suggests that it is

[1] A number of these philosophical issues are discussed in Matt Carter, *Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence* (Edinburgh: Edinburgh University Press, Ltd., 2007), 202-206.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

high time for philosophers to get into the picture and discuss the ethical implications of the rise of intelligent machines.

Ethical Machines and the Singularity

The idea of artificial moral agents (AMAs) refers to artificial intelligent systems capable of moral reasoning, judgment, and decision-making. It forms part and parcel of what Goertzel[2] defines as artificial general intelligence (AGI).[3] Computer systems with human-like general intelligence are AGIs. Contrasted with "narrow artificial intelligence" or "narrow AI," which only exhibits intelligence regarding "specific, narrowly constrained problems,"[4] AGIs exhibit a variety of human-like, intelligent behavior. Chess computer programs, for instance, could be classified as narrow AI systems because they are only considered to be intelligent in a single human domain—that is, playing chess. In contrast, an AGI is hypothesized to be intelligent in most, if not all, aspects of human cognition. Since one of the central human cognitive abilities is the capability to reason about moral issues, AGIs should, therefore, include the intellectual activity of moral reasoning (i.e., AGIs should be AMAs as well).

The idea of a singularity, on the other hand, refers to a hypothesized, futuristic event where greater-than-human-level intelligence exists—intelligences that are deemed to be a natural offshoot of modeling AI systems.[5] From this event, it is further hypothesized, that an intelligence explosion would follow.[6] If such an event were to happen, it would put into question a lot of our default notions about reality, truth, and life, including most of our ideas about what is right and wrong.

---

[2] See Ben Goertzel, "Human-level Artificial General Intelligence and the Possibility of a Technological Singularity: A Reaction to Ray Kurzweil's *The Singularity is Near*, and McDermott's Critique of Kurzweil," in *Artificial Intelligence*, 171:18 (2007), 1161-1173, doi:10.1016/j.artint.2007.10.011.

[3] Throughout the entire paper, the acronyms AI and AGI are used interchangeably.

[4] *Ibid.*, 1162.

[5] The idea of a greater-than-human-level intelligence being a consequence of developing human-level AIs is further discussed in the subsequent sections. Note that such view may be traced in Irving John Good, "Speculations Concerning the First Ultraintelligent Machine," in *Advances in Computers*, vol. 6, ed. by Franz L. Alt and Morris Rubinoff (New York: Academic Press, 1966). In addition, the same view is also discussed in David J. Chalmers, "The Singularity: A Philosophical Analysis," in *Journal of Consciousness Studies*, 17:9-10 (2010), 7-65.

[6] See Vernon Vinge, "The Coming of Technological Singularity: How to Survive in the Post-Human Era," in *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace – Proceedings of A Symposium Cosponsored by NASA Lewis Research Center and the Ohio Aerospace Institute* (Washington, D.C.: National Aeronautics and Space Administration, Office of Management, Scientific and Technical Information Program, 1993), 11-22.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

The term "singularity" has long been used in mathematics and (astro)physics. In mathematics, the singularity refers to "a value that transcends any finite limitation."[7] For example, the function $y = 1/x$ leads to a singularity since the value of y increases as x approaches zero (e.g., 1/0.0001 is greater than 1/0.001 as these operations result to 10,000 and 1,000, respectively). Keeping in mind that any number divided by zero yields a mathematically undefined result, the quotient reaches infinity as the value of x approaches this mark.[8] In physics, the notion of a singularity could be best understood by looking at the aftermath of a star's death (i.e., when a supernova occurs). Kurzweil explains that a "singularity" is created at the center of a massive star that undergoes a supernova explosion.[9] The remnant of the star collapses to this center, which is said to be a point of apparently zero volume and infinite density. This rupture in the space-time fabric is then called a black hole.[10]

The basic idea of a singularity within the fields of mathematics and physics, therefore, highlights instances wherein our standard models for understanding things just breakdown. This means that our most up-to-date theories cannot account for the phenomenon that needs explaining. For the present study, we consider another type of singularity—that is, the posited (technological) singularity that centers on intelligence.

Science fiction literature serves as a storehouse of scientific knowledge, and it is able to project future technologies that are not yet available today.[11] In a way, the same may be said with regard to the

---

[7] Ray Kurzweil, *The Singularity is Near: When Humans Transcend Biology* (New York: Viking, 2005), 35-36.

[8] Note that there are mathematicians, like Nicholas of Cusa, who consider the concept of infinity as unknowable. See Jean-Michel Counet, "Mathematics and the Divine in Nicholas of Cusa," in *Mathematics and the Divine: A Historical Study*, ed. by Teun Koetsier and Luc Bergmans (Amsterdam: Elsevier B.V., 2005.), 271-290.

[9] Kurzweil, *The Singularity is Near*, 36.

[10] The difficulties of studying the very nature of black holes are well documented. For instance, see Brian Greene, *The Elegant Universe: Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory* (New York: W.W. Norton & Co., 1999). Here it is highlighted that, in physics, it is widely accepted that Einstein's theory of general relativity best fits the study of huge and heavy things (e.g., stars, planets, etc.). Quantum mechanics, on the other hand, is used to deal with small and light objects such as atoms, subatomic particles, and so on. The problem with black holes is that, as defined above, it is both heavy and small. So, does one use general relativity because black holes have an infinite density? Or, should quantum mechanics be used instead given that it is a finite point? Furthermore, all the laws (of physics) supposedly break down in such singularities. This idea is also highlighted in Stephen W. Hawking, *A Brief History of Time* (New York: Bantam Books, 1988) and Roger Penrose, "Black Holes," in *The World Treasury of Physics, Astronomy and Mathematics*, ed. by Timothy Ferris (New York: Little, Brown and Company, 1991).

[11] This idea is discussed in Walter Moser and Craig Moyes, "Literature—A Storehouse of Knowledge?" in *SubStance*, 22 (1993), 126-140.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

development of machines with artificial minds. Carter, for instance, notes that the abstract idea of "mechanical men" or robots may be traced back to the science fiction literature of the early to mid-twentieth century.[12]

The idea of robots, or inanimate objects coming to life, dates back to the early Greeks, especially in the story of the mythological god Hephaestus, who created sophisticated machines.[13] The term "robot," on the other hand, was first used in Čapek's play, *R.U.R.*,[14] to refer to human-like creatures capable of intelligence. In a sense, before there was a full-blown academic discipline devoted to the study of both robotics and artificial intelligence, science fiction literature toyed with this idea first. The same turn from fiction to reality is envisaged by futurists concerning the idea of an intelligence explosion or a singularity.[15]

In the discipline of artificial intelligence, the term "singularity" refers to a point in human history that would drastically change life on earth as this would mark the creation of greater-than-human intelligence, followed by an intelligence explosion.[16] To differentiate this from its counterpart concepts in mathematics and astrophysics, this has been often dubbed as the "technological singularity."[17]

Many credit the statistician and computer scientist I.J. Good as one of the pioneers of the idea of a singularity, and the first who thought of the possible implications of an intelligence explosion through the rise of intelligent machines.[18] The main idea is that, after the creation of the first ultraintelligent machine (i.e., a greater-than-human artificial intelligence system), an intelligent explosion would occur, since the first AI would (mass) produce the next generation of higher intelligent machines. In effect, by iterating this process, the result would be the creation of a whole assembly line of intelligent machines.[19] In Good's own words:

---

[12] See Carter, *Minds and Computers*, 1.

[13] See Stefanos A. Paipetis, *The Unknown Technology in Homer* (Dordrecht: Springer Science+Business Media BV, 2010), 107-111.

[14] See Karel Čapek, *R.U.R.*, trans. by David Wyllie (Adelaide: eBooks@Adelaide, University of Adelaide Library, 2016).

[15] Moravec's idea of robot intelligence surpassing human intelligence before the year 2050 is reminiscent here. See Hans P. Moravec, "Rise of the Robots," in *Understanding Artificial Intelligence*, ed. by Sandy Fritz (New York: Warner Books, Inc., 2002), 114.

[16] See Kurzweil, *The Singularity is Near*, 24-25.

[17] For the purposes of this work, the term "singularity" would be used to simply refer to this.

[18] See Richard Loosemore and Ben Goertzel, "Why an Intelligence Explosion is Probable," in *Humanity+ Magazine* (7 March 2011), <http://hplusmagazine.com/2011/03/07/why-an-intelligence-explosion-is-probable/>.

[19] As discussed earlier, this is somewhat comparable to the value of x that becomes larger and larger. See Kurzweil, *The Singularity is Near*, 35-36.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. … Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously.[20]

Good argues that a direct consequence of developing ultraintelligent machines is the creation of more sophisticated AIs. These further creations, however, would not be designed by human beings any longer. For him, it is hard not to think of the scenario that these machines would eventually perform the same creative endeavors of modeling intelligent systems. But, given that this first generation of ultraintelligent machines are more intelligent than humans, it is conceivable that they would also be able to develop more sophisticated AI machines that could further surpass their own intellectual capabilities. In relation to this, also explaining the potential reason why these AI systems might undergo such process of self-improvement, Muehlhauser and Bostrom argue that:

> We can predict that advanced AIs will have instrumental goals to preserve themselves, acquire resources, and self-improve, because those goals are useful intermediaries to the achievement of almost any set of final goals. Thus, when we build an AI that is as skilled as we are at the task of designing AI systems, we may thereby initiate a rapid, AI-motivated cascade of self-improvement cycles. Now when the AI improves itself, it improves the intelligence that does the improving, quickly leaving the human level of intelligence far behind.[21]

---

[20] Good, "Speculations Concerning the First Ultraintelligent Machine," 33.

[21] Luke Muehlhauser and Nick Bostrom, "Why We Need Friendly AI," in *Think*, 36:13 (2014), 42.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

This is why Good states that the first ultraintelligent machine would arguably be the last invention of humankind. It is because all future generations of AI would be made by their fellow machines and so on. One other philosopher who has recently scrutinized the idea of the singularity is David Chalmers. For him, there are pros and cons to be considered.

For Chalmers, the singularity could be considered one of the most significant events in human history.[22] In a post-singularity world, life on earth may drastically change to the point of human incomprehensibility. Consider the impacts (i.e., both the positive and negative ones) that could result from the creation of sophisticated AI systems.

For the positive effects of a singularity, it is very likely that remedies for certain diseases deemed incurable at present (e.g., HIV virus, cancer, etc.) could be discovered eventually as the intelligence level of those on earth continuously grows. Several social problems such as racial discrimination, food scarcity, and poverty, among others, could also be addressed by such intelligent beings, not to mention resolving age-old mathematical, scientific, and even philosophical puzzles, to name a few. Perhaps, it might be the case that these concerns only pervade us today because the current intellectual capabilities of humans are quite limited. The kinds of beings that would exist after the singularity, in contrast, would be of greater-than-human intelligence. Thus, it could be held that these problems could be solved by them.[23] However, even though a singularity could yield positive outcomes, it could also produce negative ones.

There are a number of potential dangers that might result from an intelligence explosion. Among these include the "end to the human race, an arms race of warring machines, [and] the power to destroy the planet."[24] The annihilation of humanity is not as far-fetched as one may think given that it is possible that AI systems and their next generations could have a different set of values compared to human beings.[25] To hammer more on this point, consider a rough analogy motivated by the question: "Could it really be explained to pigs (i.e., making them fully understand) why it is ethically acceptable to slaughter them?"

It could be said that pig's meat is staple food for us, humans, and that this is a reasonable way of justifying why we slaughter and eat them. However, it might be difficult, if not impossible, to make pigs really understand this reason. Setting aside the obvious problem of conversing with

[22] See Chalmers, "The Singularity: A Philosophical Analysis," 4.

[23] Compare this with Kurzweil's view that any problem, insoluble as it may seem at present, has a corresponding solution (i.e., an idea). See Kurzweil, *The Singularity is Near*, 23-25.

[24] Chalmers, "The Singularity: A Philosophical Analysis," 4.

[25] See *ibid*., 24-29.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

one another, it could be said that, given their intellectual capacities, they would not be able to fully comprehend such rationale.[26]

By the same token, it is plausible to suppose that the projected AI systems in the future would have a different value system from us given their higher level of intelligence. To further explain this, consider the different ways of arriving at artificial intelligence, which would, ideally, have to be considerate of human values as well. Chalmers, for instance, identifies two options, which are the human-based approaches and non-human-based ones.[27]

Methods for Modeling Artificial Intelligence

For Chalmers, there are two ways to develop artificial general intelligence, namely: human-based and non-human-based methods.[28] Tracks toward the creation of artificial intelligence that employ human-based options extend or upgrade the biological makeup of humans.[29] This means that such methods aim to duplicate or simulate the biological brains of human beings.[30] The advantage of this approach is that "[t]he resulting systems are likely to have the same basic values as their human sources."[31] The ethical values of the ensuing AI creations that employ the human-based path would remain unchanged (i.e., to what they originally had before their enhancement procedure).

Non-human-based options, on the other hand, build AIs through designing computer programs, learning systems, or any other means that does not enhance the biological constituents of humans. For these methods, Wallach and Allen identify three possible ways,[32] namely: the top-down or direct programming track, bottom-up or developmental approaches, and the hybrid of these two.

---

[26] Here, we could also relate Wittgenstein's idea that, even if a lion could speak, human beings would not be able to understand it. See Ludwig Wittgenstein, *Philosophical Investigations* (Oxford: Blackwell Publishing, Ltd., 1953), 190.

[27] See Chalmers, "The Singularity: A Philosophical Analysis," 25-27.

[28] For the purposes of this paper, the human-based and non-human based options will be discussed in light of how they may account for ethical AI systems.

[29] See Chalmers, "The Singularity: A Philosophical Analysis," 25-27.

[30] This coincides with Vinge's four tracks towards creating greater-than-human-intelligence. See Vinge, "The Coming of Technological Singularity: How to Survive in the Post-Human Era." Note that Vinge's intelligence amplification and biological approaches share certain commonalities with Chalmers' concept of human-based artificial intelligence. This is because both support the idea of creating intelligent systems via harnessing the physiological makeup of human beings.

[31] Chalmers, "The Singularity: A Philosophical Analysis," 25.

[32] An overview of all these tracks are highlighted in Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (New York: Oxford University Press, 2009).

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

Top-down approaches subscribe to the idea that artificial moral agents may be developed via programming moral principles into artifacts so that their actions and behaviors would be regulated by such precepts. Supporters of this track claim that human thinking processes resemble the ones implemented by computational systems. Thus, they argue that intelligent behavior may be instantiated by creating computer programs. Bottom-up options, on the other hand, employ evolutionary, learning, or developmental methodologies. Such approach enables machines to learn ethically-related concepts, for instance, via interacting with other agents in their respective environments. In a way, the manner by which AI systems learn and develop is similar to how children undergo the socialization process—that is, through learning things from scratch. Integrating the design principles of both top-down and bottom-up tracks is what is known as the hybrid option.

The general idea behind the hybrid method is to embed autonomous artifacts with certain moral theories and, at the same time, allow it to further develop such principles, if not learn and acquire new ones, through the process of interacting with other agents. For Wallach and Allen, a feasible way of building AIs via the hybrid method is through the use of a connectionist model.[33] Note that, for connectionists, the human cognitive architecture is just a series of networks that employs parallel distributed processing. Thus, advocates of connectionism emulate human brains via developing artificial neural networks.[34]

The top-down, bottom-up, and hybrid approaches are some of the ways of constructing non-human-based artificial general intelligence. Note that, in contrast to human-based methods, such non-human-based tracks do not upgrade or make use of the biological makeup of humans. Thus, with regard to requiring autonomous agents to be moral agents as well, human-based approaches supposedly have the upper hand.[35] This is because the ethical values of the resulting AIs from human-based methods would remain unchanged, which further implies that there is a higher risk of building advanced AIs via the non-human-based route. This is not to say, however,

[33] How a connectionist model may be considered a hybrid model is discussed in length by Wallach and Allen. See *ibid*., 121-124.

[34] It is believed that artificial neural networks could integrate the design principles of top-down and bottom-up routes. This is because such networks have two separate kinds of connection weights, namely: hardwired and learned connection weights. Supposedly, these weights may be considered to share the same principles with the top-down and bottom-up paths, respectively. For a discussion on connectionism, see Napoleon Mabaquiao, Jr., *Mind, Science and Computation* (Manila: Vibal Publishing House, Inc., 2012), 124-129.

[35] Chalmers further argues that human-based approaches are not "extendable," and, as such, might not lead to a singularity. See Chalmers, "The Singularity: A Philosophical Analysis," 26.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

that the human-based path is foolproof. It could be argued that there is no certainty as to what would happen to their set of values once they also upgrade their computational, or thinking, capacities.[36]

The idea of artificially intelligent machines having a different value system from humans could somehow be drawn from the design methodologies mentioned above. Unfortunately, for us, humans would be the lesser beings as compared to such superior intelligence. If a singularity were to happen, then human values would likely be incommensurable with these AI systems. Thus, the extinction of humankind may be foreseeable if safety measures would not be considered. With the risks involved in creating artificial intelligence, the question is, what now is the job of the philosopher in all of these issues?

Technical and Foundational Problems

Research on artificial intelligence has been encountering a number of problems for many years now. Such obstacles could be classified as either technical or foundational ones. Philosophers who wish to contribute in addressing, if not preventing, the potential dangers of a singularity may do so by doing research on the latter.

Delineating one from the other, the technical problems of artificial intelligence may be defined as those that are presently encountered by the field given the current state of technology. However, once certain technological advancements occur, these issues, in theory, would cease to exist. Among these technical problems include issues regarding robustness and generalization, real-time processing, and the sequential nature of programs, among others.[37]

For example, the issue on robustness and generalization focuses on the problem of noise and fault tolerance. Additionally, it also includes the issue regarding the capability of artifacts to act and react aptly in novel situations. Note that noise is defined as random data fluctuations. So, an artifact is said to be noise-tolerant if it could still process data that contains fluctuations. On the other hand, if an artifact performs adequately in spite of its faulty components, then it is considered to be fault-tolerant.

Analogous to the issue of robustness, autonomous machines often lack the capability of acting and reacting to novel situations that were not originally programmed into them by their designers. This is because they lack generalization capabilities, which eventually results to an artifact halting or

[36] In relation to this, see the same article of Chalmers, specifically his discussion on Kant's view regarding values and rationality. See *ibid*.

[37] Rolf Pfeifer and Christian Scheier, *Understanding Intelligence* (Cambridge: MIT Press, 1999), 63-64.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

breaking down. However, the issue of both robustness and generalization may be considered technical ones as they are solvable, in principle, at the advent of technological improvements. In the future, once the current state of technology progresses, these technical problems would foreseeably go away.

As for the foundational issues encountered by artificial intelligence research, these problems are considered philosophical in nature as they question the very foundations, or the foundational assumptions, of the methodologies themselves. Standard examples of these are the symbol-grounding problem and frame problem.

Pfeifer and Scheier explain that the symbol-grounding problem centers on the issue of how the closed system of physical symbols employed by classical artificial intelligent machines relate to the actual world.[38] Note that physical symbols are defined purely in a syntactical manner. They follow the law of representation, which states that situations in the real world may be mapped into internal representations. So, these physical symbols are processed by a central processing system solely based on their syntax. The question is this: "If physical symbols correspond to specific things in the real world, how then could a syntax-based system ground and derive their meanings?" Put in another way, if traditional AI systems process symbols in a purely syntactical manner, then deriving or inferring the meanings of the employed symbols would still not have been accounted for.

On the other hand, the frame problem focuses on the interaction between a modeled system and its corresponding environment. Simply put, this philosophical problem focuses on "how models of a changing environment can be kept in tune with the environment."[39] Given that artificial intelligent systems developed via the classical artificial intelligence path, for instance, have internal programs of the real world,[40] the frame problem deals with how to properly identify from such world model those data that needs updating after an action takes place.[41] Apparently, the task of keeping programs of changing environments in tune with the real world presents a number of difficulties.

Note that the problems mentioned above are considered philosophical in nature as they question the foundational suppositions of the design strategies towards building AGIs. Another foundational issue that may interest philosophers deals with the problem of making artificially

---

[38] *Ibid*., 69-71.

[39] *Ibid*., 68.

[40] Note that there are those who argue that the frame problem also arises in connectionist models. For instance, see Murray Shanahan, "The Frame Problem," in *The Stanford Encyclopedia of Philosophy*, Winter 2009 ed., ed. by Edward N. Zalta, <https://plato.stanford.edu/archives/win2009/entries/frame-problem/> (July 2013).

[41] *Ibid*.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

intelligent machines capable of moral reasoning, judgment, and decision-making—that is, the creation of artificial moral agents. As discussed earlier, the extinction of human beings may be a likely consequence if safety measures are not considered in designing AI systems. This is the reason why the research area of machine ethics looks into the possibility of modeling the first batch of AIs as being considerate of human values as well.

Machine Ethics and the Nature of Moral Agents

Michael Anderson and Susan Anderson tell us that:

> [M]achine ethics is concerned with giving machines ethical principles or a procedure for discovering a way to resolve the ethical dilemmas they might encounter, enabling them to function in an ethically responsible manner through their own ethical decision making.[42]

What this tells us is simple: machine ethics is concerned about the creation of AMAs—artificial intelligent machines capable of moral reasoning. Note that this field is often contrasted with the closely related discipline of technology ethics. In a nutshell, technology ethics is a branch of applied ethics largely devoted in developing ethics for human beings who use machines or technology.[43]

Distinguishing machine ethics from the philosophy of technology is important, since the latter is more concerned with the ethical standing of human beings who use technological products such as intelligent machines (i.e., it looks at the proper and improper human behavior with regard to the usage of machines). Thus, it considers machines as tools and not as autonomous agents. Machine ethics, in contrast, regards machines as actual or potential moral agents.

For machine ethicists, moral praise and blame could be attributed to the actions of autonomous agents, and it seems that there are good reasons to think that, indeed, sophisticated technologies may be considered as moral agents. Note that moral agents are specific kinds of entities whose behaviors are subject to moral requirements (i.e., under a set of ethical standards, moral praise or blame could be ascribed to its actions).[44]

---

[42] Michael Anderson and Susan Leigh Anderson, *Machine Ethics* (New York: Cambridge University Press, 2011), 1.

[43] See Wallach and Allen, *Moral Machines*, 37-39.

[44] See Kenneth Einar Himma, "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties must an Artificial Agent have to be a Moral Agent?" in *Ethics and Information Technology*, 11 (2009), 21.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

In the philosophical tradition, discussions regarding the nature of moral agents are nothing new. Several philosophers have already put forward their own take on the nature of moral agents. For example, Aristotle:

> [i]n the course of discussing human virtues and their corresponding vices … begins with a brief statement of the concept of moral responsibility—that it is sometimes appropriate to respond to an agent with praise or blame on the basis of her actions and/or dispositional traits of character. … A bit later, he clarifies that only a certain kind of agent qualifies as a moral agent and is thus properly subject to ascriptions of responsibility, namely, one who possess a capacity for decision. For Aristotle, a decision is a particular kind of desire resulting from deliberation, one that expresses the agent's conception of what is good.[45]

In a way, such notion parallels Aquinas' idea that "[g]ood ends and means are those befitting the human agent."[46] For Aquinas, as with Aristotle, the said agent should be capable of deliberating and judging what action is good,[47] and these actions are called human actions—that is, actions that which one has voluntary control of or those that result from certain free judgements. To account for the nature of such free judgments, Aquinas explains:

> The proper act of free-will is choice: for we say that we have a free-will because we can take one thing while refusing another; and this is to choose. Therefore we must consider the nature of free-will, by considering the nature of choice. Now two things concur in choice: one on the part of the cognitive power, the other on the part of the appetitive power. On the part of the cognitive power, counsel is required, by which we judge one thing to be preferred to another: and on the part of the

[45] See Andrew Eshleman, "Moral Responsibility," in *The Stanford Encyclopedia of Philosophy*, Summer 2014 ed., ed. by Edward N. Zalta, <https://plato.stanford.edu/archives/sum2014/entries/moral-responsibility/> (June 2015).

[46] Ralph McInerny, "Aquinas's Moral Theory," in *Journal of Medical Ethics*, 13:1 (1987), 31-33.

[47] Thomas Aquinas, *Summa Theologica*, trans. by the Fathers of the English Dominican Province, rev. by Daniel J. Sullivan (Chicago: Encyclopaedia Britannica, 1952), IaIIae 1.1.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

> appetitive power, it is required that the appetite should accept the judgment of counsel.[48]

So, for Aquinas, moral agents are those who have mastery over one's actions, exemplified by their ability to prefer what is good. This same line of thinking seems to be echoed by other philosophers such as Kant. For Kant, "a moral agent is autonomous in that it both gives itself the moral law (it is self-legislating) and can constrain or motivate itself to follow the law (it is self-constraining or self-motivating)."[49] In such characterization, note that Kant highlights the notion of a moral agent being autonomous.[50] To a certain extent, these characterizations of moral agency, along with the other philosophical theories[51] that try to account for its nature, relate to the idea of building artificial moral agents.

Safeguarding Humanity via Artificial Moral Agents

One possible way of preventing the negative outcomes of a singularity would be in terms of how philosophers could further involve themselves in addressing problems related to artificial moral agency. For one, there is the issue regarding the nature of such agents.

In terms of accounting for artificial moral agency, several theories have been proposed by those working under the field of machine ethics. Sullins,[52] for instance, defines artificial moral agents as artificial autonomous agents[53] that possess moral value. He first explains that:

---

[48] *Ibid*., Ia 83.3.

[49] Lara Denis, "Kant and Hume on Morality," in *The Stanford Encyclopedia of Philosophy*, Fall 2012 ed., ed. by Edward N. Zalta, <https://plato.stanford.edu/archives/fall2012/entries/kant-hume-morality/> (7 December 2015).

[50] For Kant's discussion regarding the nature of moral agents, specifically the concepts of will and autonomy, see Immanuel Kant, *Groundwork for the Metaphysic of Morals*, ed. and trans. by Allen W. Wood (Yale University Press, 2002), 4:440.

[51] Other theories on moral agency are discussed extensively in Theodore C. Denise, *Great Traditions in Ethics* (California: Thomson Wadsworth, 2008).

[52] See John Sullins, "Artificial Moral Agency in Technoethics," in *Handbook of Research on Technoethics*, ed. by Roccio Luppicini and Rebecca Adell (Hershey: IGI Global Information Science, 2009), 205-221.

[53] In defining autonomous agents as systems that act upon their situated environment in pursuit of their own agenda, Sullins cites Stan Franklin and Art Graesser, "Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents," in *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages* (London: Springer-Verlag, 1996). Such characterization implies that agents are presupposed as entities that have causal influence or effect on other agents and their environment. Thus, any piece of technology, a sophisticated robot, for instance, that acts as an agent is considered an autonomous agent.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

> [I]deally, the agent should be a continuous process that monitors its environment, be able to communicate with its user or at least other simple agents, have some sort of machine learning, and be mobile in its environment or able to move to other environments, as well as be flexible in its reactions to stimulus in these environments. It is also a bonus if the artificial agent is affective in its character in order to interact with a realistic personality that will ease communications with human agents.[54]

Note that these specific qualities are not necessary conditions for artificial agency but are conditions that may aid artificial agents in trying to interact with others and the rest of the world.

On top of these qualities, three other conditions would have to obtain in order to ascribe moral agency to artifacts. These are the conditions of autonomy, intentionality, and responsibility. For an artifact to have the capacity to exhibit moral responsibility, it should possess significant autonomy. Thus, an artifact should be capable of performing autonomous actions (i.e., it should be able to implement tasks or goals independent of any other agent). In addition, agents must also possess the capability of acting intentionally. Lastly, it would be possible to ascribe moral agency to artifacts if its behaviors would only make sense by assuming that it has responsibility towards other moral agents. These three conditions supposedly provide a deeper understanding on the nature of AMAs. For Sullins, as long as these things obtain, an artifact could be said to be an artificial moral agent.

In contrast to Sullins, Moor proposed a four-tier categorization of artifacts in terms of appraising their moral status.[55] At the bottom-most level are ethical-impact agents. These machines are evaluated based on the moral consequences they produce. Next, machines that have built-in safety features could be considered as implicit ethical agents. To promote ethical behavior, the internal mechanisms of such machines have already been designed to consider potential safety and reliability issues. Explicit ethical agents, on the other hand, are a tier higher than implicit ethical agents. This is because such machines already have some capacity to exhibit moral reasoning. At the top-most level of this hierarchy are full ethical agents. Average adult human beings are an instance of this type of agent. Artificial moral agents, for Moor, are somewhere in-between explicit ethical agents and full ethical ones.

---

[54] Sullins, "Artificial Moral Agency in Technoethics," 207.

[55] James H. Moor, "The Nature, Importance, and Difficulty of Machine Ethics," in *Machine Ethics*, ed. by Michael Anderson and Susan Leigh Anderson (New York: Cambridge University Press, 2011), 13-20.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

A third alternative is provided by Wallach and Allen. They argue that the path towards the creation of artificial moral agents is through considering the conditions of autonomy and ethical sensitivity.[56] First, they categorize machines as either having operational morality or functional morality. Machines with operational morality are those that are capable of taking into account the morally-relevant aspects of a given situation with the aid of human architects. These architects embed ethical considerations into the said machines. Machines that fall under functional morality, on the other hand, are those that possess the capacity to exhibit some form of moral reasoning and decision-making.

Such categorization forms a spectrum. One finds machines that fall under operational morality on one end of this spectrum and, on the other end, are full moral agents—that is, artifacts that have high autonomy and high ethical sensitivity. Moreover, machines that fall under functional morality are distributed in-between of these two extremes. Wallach and Allen claim that this two-dimensional framework could possibly account for artificial moral agency since the conditions of autonomy and ethical sensitivity serve as benchmarks as to what counts as a full moral agent. Any machine approaching high autonomy and high ethical sensitivity will be counted as a full moral agent, and this could be achieved by incrementally improving these conditions.

At present, there seems to be no hard consensus as to which artificial moral agency framework would likely yield a positive result. For instance, the idea of reducing the three mentioned above into a single, all-encompassing theory has not yet been fully explored. Furthermore, how exactly shall we proceed in actually realizing artificial moral agents is another problem in machine ethics. These issues, among the many others that also need final resolution, should be the business of philosophers. With the body of work that they have contributed in the study of moral agency, philosophers should continue to extend their research in the field of machine ethics.

Summary and Conclusion

As discussed above, the singularity is a hypothesized, futuristic event that pertains to the invention of machines that are of greater-than-human-level intelligence. This scenario may be considered a natural consequence of developing AI systems. Once these things are fully realized, an intelligence explosion would follow soon after. Note that such runaway would seemingly put into question humankind's standard concepts about reality, life, and so on, and this would also include our general ethical notions.

[56] Wallach and Allen, *Moral Machines*, 25-33.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

At present, there is no definitive way on how to make artificially intelligent systems value what humans do. The idea of machines exterminating the entire human race, therefore, is not, strictly speaking, a tall tale. As highlighted by both Vinge and Chalmers, there are plausible threats. To prevent these dangers, those working in artificial intelligence research, including philosophers, should take the problem of creating artificial moral agents more seriously. Philosophers, for instance, should further examine the philosophical tenability of the top-down, bottom-up, and hybrid options, among others, of modeling such agents. For one, as noted earlier, the actual nature of AMAs is still an open question. With their significant contributions in the study of moral agency, philosophers should, therefore, further examine this issue.

Other philosophical concepts such as moral reasoning, moral agency, and others may also be re-examined by philosophers as they closely relate to notion of artificial moral agency. For example, with regard to the notion of moral agency, there are those who argue that humans are not really moral agents. Nadeau, for instance, maintains that robots would be, in fact, the first moral agents to inhabit earth, if ever.[57] If we consider such view, then the question is this: if our conception of moral agency is predicated on the view that human beings are moral agents, then what happens to this notion if it is proven that the latter claim is false? In addition, there are other issues related to AI research that may also be further studied by philosophers. Among these include problems regarding the concept of consciousness and personal identity.[58]

Analyzing the foundational issues surrounding artificial intelligence research is key in safeguarding the future of humanity. Considering the gravity of the potential negative outcomes of a singularity, factor in also the open questions in modeling AMAs, it may be the case that there would not be an overabundance of philosophers working on the issues related to machine ethics—that is, at least for now.

Department of Philosophy, De La Salle University, Philippines

References

[57] See also Joseph Emile Nadeau, "Only Androids Can Be Ethical," in *Thinking about Android Epistemology*, ed. by Kenneth M. Ford, Clark Glymour and Patrick Hayes (Cambridge: MIT Press, 2006), 241-248.

[58] See Chalmers, "The Singularity: A Philosophical Analysis." For a good introduction to the issues in personal identity, see Brian Garrett, *Personal Identity and Self-Consciousness* (London: Routledge, 1998). For the ramifications to questions about human values, see J. Joven Joaquin, "Personal Identity and What Matters," in *Organon F* 24:2 (2017). 196-213.

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

Anderson, Michael and Susan Leigh Anderson, *Machine Ethics* (New York: Cambridge University Press, 2011).

Aquinas, Thomas, *Summa Theologica*, trans. by the Fathers of the English Dominican Province, rev. by Daniel J. Sullivan (Chicago: Encyclopaedia Britannica, 1952).

Čapek, Karel, *R.U.R.*, trans. by David Wyllie (Adelaide: eBooks@Adelaide, University of Adelaide Library, 2016).

Carter, Matt, *Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence* (Edinburgh: Edinburgh University Press, Ltd., 2007).

Chalmers, David J., "The Singularity: A Philosophical Analysis," in *Journal of Consciousness Studies*, 17:9-10 (2010).

Counet, Jean-Michael, "Mathematics and the Divine in Nicholas of Cusa," in *Mathematics and the Divine: A Historical Study*, ed. by Teun Koetsier and Luc Bergmans (Amsterdam: Elsevier B.V., 2005).

Denis, Lara, "Kant and Hume on Morality," in *The Stanford Encyclopedia of Philosophy*, Fall 2012 ed., ed. by Edward N. Zalta, <https://plato.stanford.edu/archives/fall2012/entries/kant-hume-morality/> (7 December 2015).

Denise, Theodore C., *Great Traditions in Ethics* (California: Thomson Wadsworth, 2008).

Eshleman, Andrew, "Moral Responsibility," in *The Stanford Encyclopedia of Philosophy*, Summer 2014 ed., ed. by Edward N. Zalta, <https://plato.stanford.edu/archives/sum2014/entries/moral-responsibility/> (June 2015).

Garrett, Brian, *Personal Identity and Self-Consciousness* (London: Routledge, 1998).

Goertzel, Ben, "Human-level Artificial General Intelligence and the Possibility of a Technological Singularity: A Reaction to Ray Kurzweil's *The Singularity is Near*, and McDermott's Critique of Kurzweil," in *Artificial Intelligence*, 171:18 (2007), doi: 10.1016/j.artint.2007.10.011.

Good, Irving John, "Speculations Concerning the First Ultraintelligent Machine," in *Advances in Computers*, vol. 6, ed. by Franz L. Alt and Morris Rubinoff (New York: Academic Press, 1966).

Graesser, Art, "Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents," in *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages* (London: Springer-Verlag, 1996).

Greene, Brian, *The Elegant Universe: Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory* (New York: W.W. Norton & Co., 1999).

Hawking, Stephen W. *A Brief History of Time* (New York: Bantam Books, 1988).

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

Himma, Kenneth Eimar, "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties must an Artificial Agent have to be a Moral Agent?" in *Ethics and Information Technology*, 11 (2009).

Joaquin, J. Joven, "Personal Identity and What Matters," in *Organon F* 24:2 (2017).

Kant, Immanuel, *Groundwork for the Metaphysic of Morals*, ed. and trans. by Allen W. Wood (Yale University Press, 2002).

Kurzweil, Ray, *The Singularity is Near: When Humans Transcend Biology* (New York: Viking, 2005).

Loosemore, Richard and Ben Goertzel, "Why an Intelligence Explosion is Probable," in *Humanity+ Magazine* (7 March 2011), <http://hplusmagazine.com/2011/03/07/why-an-intelligence-explosion-is-probable/>.

Mabaquiao, Napoleon Jr., *Mind, Science and Computation* (Manila: Vibal Publishing House, Inc., 2012).

McInerny, Ralph, "Aquinas's Moral Theory," in *Journal of Medical Ethics*, 13:1 (1987).

Moor, James H., "The Nature, Importance, and Difficulty of Machine Ethics," in *Machine Ethics*, ed. by Michael Anderson and Susan Leigh Anderson (New York: Cambridge University Press, 2011).

Moravec, Hans P., "Rise of the Robots," in *Understanding Artificial Intelligence*, ed. by Sandy Fritz (New York: Warner Books, Inc., 2002).

Moser, Walter and Craig Moyes, "Literature—A Storehouse of Knowledge?" in *SubStance*, 22 (1993).

Muehlhauser, Luke and Nick Bostrom, "Why We Need Friendly AI," in *Think*, 36:13 (2014).

Nadeau, Joseph Emile, "Only Androids Can Be Ethical," in *Thinking about Android Epistemology*, ed. by Kenneth M. Ford, Clark Glymour and Patrick Hayes (Cambridge: MIT Press, 2006).

Paipetis, Stefanos A., *The Unknown Technology in Homer* (Dordrecht: Springer Science+Business Media BV, 2010).

Penrose, Roger, "Black Holes," in *The World Treasury of Physics, Astronomy and Mathematics*, ed. by Timothy Ferris (New York: Little, Brown and Company, 1991).

Shanahan, Murray, "The Frame Problem," in *The Stanford Encyclopedia of Philosophy*, Winter 2009 ed., ed. by Edward N. Zalta, <https://plato.stanford.edu/archives/win2009/entries/frame-problem/> (July 2013).

Sullins, John, "Artificial Moral Agency in Technoethics," in *Handbook of Research on Technoethics*, ed. by Roccio Luppicini and Rebecca Adell (Hershey: IGI Global Information Science, 2009).

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330

Pfeifer, Rolf and Christian Scheier, *Understanding Intelligence* (Cambridge: MIT Press, 1999).

Vinge, Vernon, "The Coming of Technological Singularity: How to Survive in the Post-Human Era," in *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace – Proceedings of A Symposium Cosponsored by NASA Lewis Research Center and the Ohio Aerospace Institute* (Washington, D.C.: National Aeronautics and Space Administration, Office of Management, Scientific and Technical Information Program, 1993).

Wallach, Wendell and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (New York: Oxford University Press, 2009).

Wittgenstein, Ludwig, *Philosophical Investigations* (Oxford: Blackwell Publishing, Ltd., 1953).

https://www.kritike.org/journal/issue_22/boyles_june2018.pdf
ISSN 1908-7330