philosophical signposts for artificial moral agent frameworks

Robert James M. Boyles
De La Salle University, Manila

Abstract

This article focuses on a particular issue under machine ethics—that is, the nature of Artificial Moral Agents. Machine ethics is a branch of artificial intelligence that looks into the moral status of artificial agents. Artificial moral agents, on the other hand, are artificial autonomous agents that possess moral value, as well as certain rights and responsibilities. This paper demonstrates that attempts to fully develop a theory that could possibly account for the nature of Artificial Moral Agents may consider certain philosophical ideas, like the standard characterizations of agency, rational agency, moral agency, and artificial agency. At the very least, the said philosophical concepts may be treated as signposts for further research on how to truly account for the nature of Artificial Moral Agents.

Keywords

machine ethics, artificial moral agent, moral agency, artificial agent, rational agency

Suri § Volume 6 No. 2 (2017): 92 - 109
Author's Correspondence to: <email>

http://suri.pap73.org/issue9/Boyles_SURI_2017.pdf

Introduction

We live in a technologically rich world. Most all of our daily activities are aided by products of technology. Think of mobile phones, laptops, and even ordinary bread toasters, to name a few. As we move towards a fully automated world where artificial machines are capable of thinking and acting for themselves, important ethical questions arise. For instance, consider the case of driverless train systems.

In September 22, 2006, it was reported that twenty-three people died in northern Germany when a driverless magnetic levitation train crashed into a maintenance truck.[1] As per the accounts of the survivors, they could only watch the train crashing into the maintenance vehicle given that there was no driver to alert. In this scenario, who should be held responsible? Is the artificial intelligent system accountable for this incident? Or should the human designers who developed the train's computer program be blamed for the said incident?

The case cited above, as well as other related ones, are those currently being studied in machine ethics, an emerging research area in the field of artificial intelligence. This article focuses on a particular issue under machine ethics—that is, the nature of Artificial Moral Agents (AMAs). It demonstrates that attempts to develop a theory that could possibly account for the latter may consider certain philosophical ideas, like the standard characterizations of agency, moral agency, rational agency, and so on.

Machine Ethics and Technology Ethics

Machine ethics, also called machine morality, artificial morality, or computational ethics, is aimed at developing artificial intelligent systems that behave ethically.[2] The motivation of this subfield is to develop artificial moral agents that are sensitive to and considerate of human values so that humanity's welfare and

[1] Luke Harding, "At least 23 Die as Driverless Train Crashes into Maintenance Truck," *The Guardian*, September 23, 2006. http://www.theguardian.com/world/2006/sep/23/germany.topstories3 (accessed May 7, 2015).

[2] Carl Shulman, Henrik Jonsson and Nick Tarleton, "Machine Ethics and Superintelligence," in *Proceedings of the AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference*, ed. Carson Reynolds and Alvaro Cassinelli (Tokyo: AP-CAP, 2009), 1.

future would be safeguarded.[3] Anderson and Anderson explain the goal and challenges of machine ethics as follows:

> The ultimate goal of machine ethics, we believe, is to create a machine that itself follows an ideal ethical principle or set of principles; that is to say, it is guided by this principle or these principles in decisions it makes about possible courses of action it could take... One needs to turn to the branch of philosophy that is concerned with ethics for insight into what is considered to be ethically acceptable behavior. It is a considerable challenge because, even among experts, ethics has not been completely codified. It is a field that is still evolving.[4]

The field of machine ethics is often contrasted with the closely related discipline of technology ethics. Technology ethics, otherwise known as philosophy of technology, is a branch of applied ethics that focuses on the development of ethics for humans who utilize machines or technology.[5] Furthermore, the area

> ...is highly interdisciplinary... [and] taken as a whole is an understanding of the consequences of technological impacts relating to the environment, the society, and human existence."[6]

To further differentiate machine ethics from the philosophy of technology, note that the latter largely deals with the ethical standing of humans who utilize technology. This means that it looks at the proper and improper human behavior

[3] This closely resembles Yudkowsky's concept of "Friendly AI." For his full discussion of the motivations behind the creation of such intelligent machines, including the potential design features and cognitive architectures, see Eliezer Yudkowsky, "Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures," Machine Intelligence Research Institute, 2001. http://intelligence.org/files/CFAI.pdf (accessed June 2, 2014).

[4] Michael Anderson and Susan Leigh Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent," *AI Magazine* 28, no. 4 (2007): 15. http://www.aaai.org/ojs/index.php/aimagazine/article/download/2065/2052 (accessed July 1, 2015).

[5] Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right From Wrong* (New York: Oxford University Press, 2009), 37-39.

[6] Jan Kyrre Berg, Olsen, Stig Andur Pedersen, and Vincent F. Hendricks, ed., *Blackwell Companions to Philosophy: A Companion to the Philosophy of Technology* (West Sussex: Blackwell Publishing Ltd., 2009), 1.

in terms of, say, using machines, which also entails that such machines are deemed as mere tools and not as autonomous agents. In contrast, machine ethics focuses on the moral status of intelligent machines. It considers these systems as actual or potential moral agents, which moral praise and blame could be attributed.

One worry that motivates theorists in machine ethics is the thought that artificial agents would be incapable of recognizing what human beings value (e.g., respect for life, freedom, etc.), which means that they could be potential threats to human existence. Bostrom and Yudkowsky,[7] for instance, have argued that this scenario is foreseeable once artificial agents perform social functions such as responsibility, transparency, and so on. This is the reason why machine ethicists inquire about the very nature of artificial moral agents.

Accounting for Artificial Moral Agency

Some AI theorists characterize artificial moral agents as artificial autonomous agents that possess moral value, as well as certain rights and responsibilities.[8] With regards to the nature of artificial moral agency, however, theorists are somehow divided on how to account for this concept. For one, a number of models have been proffered by those working under the field of machine ethics. In this article, three theories that try to explain the notion of artificial moral agency are examined, namely: Sullin's, Moor's, and Wallach and Allen's.

A contemporary AI theorist, Sullins[9] looked into three specific aspects of autonomous robots in terms of evaluating their moral status, which are autonomy, intentionality, and responsibility. These three conditions would supposedly enable moral agency ascriptions to autonomous artifacts, specifically artificial agents, like

[7] Nick Bostrom and Eliezer Yudkowsky, "The Ethics of Artificial Intelligence," in *The Cambridge Handbook of Artificial Intelligence*, ed. Keith Frankish and William Ramsey (New York: Cambridge University Press, 2013), 1-2. http://intelligence.org/files/EthicsofAI.pdf (accessed July 28, 2014).

[8] For example, see John P. Sullins, "When Is a Robot a Moral Agent? International Review of Information," *Ethics* 6 (2006): 23-30. http://www.i-r-i-e.net/inhalt/006/006_Sullins.pdf (accessed June 2, 2014) and John P. Sullins, "Artificial Moral Agency in Technoethics," in *Handbook of Research on Technoethics*, ed. Rocci Luppicini and Rebecca Adell (Hershey: IGI Global Information Science, 2009), 205-221.

[9] John P. Sullins, "When Is a Robot a Moral Agent? International Review of Information," 23-30.

robots.[10] Note that autonomous agents may be understood as rational systems that act upon their situated environment in pursuit of their own agenda,[11] which means that such agents are considered entities that have causal influence on other agents and their environment.

For Sullins, an artifact should possess autonomy for it to exhibit moral responsibility. The term "autonomy" is understood here to mean the absence of external causes. In the context of artificial agents, this implies that a robot is not directly controlled by any other agent or user. This characterization is commonly used in the field of engineering. Furthermore, as long as a robot is able to implement its goals or tasks independent of any other agent, it is capable of performing autonomous actions, which entails that it has effective autonomy.

Secondly, for a robot to be considered as a moral agent, such autonomous machine must also be capable of acting intentionally. Note here that there is no need to prove that a robot actually possesses intentionality (i.e., in the strongest sense of the term). Since this issue is very problematic even for human beings, it should not be an issue for artifacts as well. So, by the principle of fairness, it should not also be demanded that we prove conclusively that robots have intentional states. In Sullins' own words:

> There is no requirement that the actions [of autonomous robots] really are intentional in a philosophically rigorous way, nor that the actions are derived from a will that is free on all levels of abstraction. All that is needed is that, at the level of the interaction between the agents involved, there is a comparable level of personal intentionality and free will between all the agents involved.[12]

Finally, Sullins maintains that moral agency ascriptions to autonomous artifacts is possible if its behaviors would only make sense by assuming that it has

[10] For some information about how the concept of moral responsibility specifically relates to technological artifacts see Merel Noorman, "Computing and Moral Responsibility," Stanford Encyclopedia of Philosophy, ed. Edward Zalta, 2012. http://plato.stanford.edu/entries/computing-responsibility (accessed June 2, 2015).

[11] 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).

[12] John P. Sullins, "When Is a Robot a Moral Agent? International Review of Information," 26.

responsibility to other moral agents. For instance, if a fairly autonomous robot is given a certain societal role, this means that it should also be cautious of the specific responsibilities that come along with the said function. So, it could be assumed that a robotic caregiver takes into consideration and is mindful of the well-being of its "patients" every time that it performs effectively. Such a kind of (ethical) action, therefore, could only be explained if one assumes that the artifact really understands its responsibilities with regards to the whole health care system. In one sense, it could be said that this line of reasoning is an argument from the best possible explanation (i.e., the seemingly ethical behavior of a robot could only be explained if one presumes that it already understands its own moral obligations).

The point of underscoring the three conditions proposed by Sullins is that it provides a deeper understanding of the idea that AMAs are artificial autonomous agents that embody moral value, rights, and responsibilities. As long as these conditions obtain, an artifact could be said to be an artificial moral agent.

On the other hand, another AI theorist, Moor,[13] has offered a four-tier categorization of artifacts in terms of appraising their moral status: ethical-impact agents, implicit ethical agents, explicit ethical agents, and full ethical agents.

According to Moor, ethical-impact agents are at the bottom-most level.[14] Such machines are generally evaluated based on the moral consequences they produce. Given this characterization, it could be said that land mine detecting robots, like the modified Husky UGV robot of the University of Coimbra, Portugal are ethical-impact agents,[15] since the creation of such robots generally produce a good (moral) outcome by lessening the loss of lives of human minesweepers.

Secondly, machines that have built-in safety features, or were especially designed so that negative ethical effects may be avoided, are what Moor considers as implicit ethical agents.[16] Examples of these include automated teller machines, auto-piloted planes, and so on. All these are implicit ethical agents because their

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

[14] *Ibid*., 15.

[15] Evan Ackerman, "Robot Takes on Landmine Detection While Humans Stay Very Very Far Away," *IEEE Spectrum*, January 23, 2014. http://spectrum.ieee.org/automaton/robotics/military-robots/husky-robot-takes-on-landmine-detection-while-humans-stay-very-very-far-away (accessed June 25, 2015).

[16] James H. Moor, "The Nature, Importance, and Difficulty of Machine Ethics," 15-16.

actions are constrained in such a way that it averts unethical outcomes. To do this (i.e., for a machine to promote ethical behavior), the internal functions of such technologies are designed to consider potential safety and reliability issues.

Thirdly, Moor views artifacts that have the capacity to reason about ethics as explicit ethical agents.[17] These types of machines are developed by mapping out ethical categories into their internal programming. Basically, given that these artifacts are embedded with ethical precepts in their machinery, they would, in turn, be capable of making ethical judgments when faced with moral dilemmas. On top of this, they would also be able to justify such judgments. IBM's supercomputer, Watson might be a close example of this type of agent. Currently, IBM's Watson is "studying" how to diagnose medical patients in the hopes that it would aid physicians in identifying the nature of complicated diseases.[18]

Finally, on top of the hierarchy are full ethical agents. As compared to explicit ethical agents, these machines are more advanced as they are able to exhibit "explicit ethical judgments and generally [are] competent to reasonably justify them."[19] An average adult human being is said to be an example of this type of agent, and this is largely due to the idea that they have consciousness, intentionality, and free will.

The goal of machine ethics should be to create explicit ethical agents.[20] Given Moor's four-tier categorization, AMAs might be seen somewhere between explicit ethical agents and full ethical ones.

Another way of characterizing artificial moral agents is by building machines with significant autonomy and designing such things to be sensitive to morally-relevant facts as they freely interact in the real world. Such strategy would entail artifacts that possess moral responsibility. Wallach and Allen,[21] for example, contend that the pathway towards the full implementation of sophisticated AMAs is via considering the said conditions of autonomy and ethical sensitivity. For them:

---

[17] *Ibid*., 16-18.

[18] Lauren Friedman, "IBM's Watson May Soon Be The Best Doctor In The World," Business Insider, April 22, 2014. http://www.businessinsider.com/ibms-watson-may-soon-be-the-best-doctor-in-the-world-2014-4 (accessed June 25, 2015).

[19] James H. Moor, "The Nature, Importance, and Difficulty of Machine Ethics," 18.

[20] Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right From Wrong*, 34.

[21] *Ibid*., 25-39.

> …[the] framework [towards building AMAs] has two dimensions: autonomy and sensitivity to values. These dimensions are independent, as the parent of any teenager knows. Increased autonomy is not always balanced by increased sensitivity to the values of others; this is as true of technology as it is of teenagers.[22]

First, Wallach and Allen distinguish technologies that fall under the category of operational morality from those that fall under functional morality. For example, a gun that has a childproof safety mechanism falls under operational morality, since the determination of ethical values and what factors are ethically relevant to a given situation are not determined by the machines themselves; rather, were already thought out by their architects during the design process. Automobiles equipped with airbag, safety belt, and child safety lock devices fall under this taxonomy as well. Autopilot aircrafts and medical ethics expert systems, in contrast, maybe said to fall under functional morality, since these two have the capacity of exhibiting some form of moral reasoning and decision-making. Note that these are capabilities that we often associate with autonomy and ethical sensitivity.

Wallach and Allen's categorization forms a spectrum. On one end of this spectrum, we find machines that fall under operational morality. On the other end are the full moral agents (i.e., those that have high autonomy and high ethical sensitivity). Machines that have functional morality, on the other hand, maybe found between these two extremes. Some might have high autonomy but low ethical sensitivity (e.g., autopilot aircrafts). Others might have low autonomy but high ethical sensitivity, like medical ethics expert systems. Wallach and Allen claim that this two-dimensional framework could possibly aid us in developing AMAs, since they will serve as standards as to what may count as a full moral agent. Furthermore, by incrementally improving the said conditions, any sophisticated machine that is close to having high autonomy and high ethical sensitivity will be counted as a full moral agent.[23]

The proposed theories of Sullins, Moor, and Wallach and Allen are some of the different attempts that try to account for the nature of artificial moral agency. There is no hard consensus, however, on which candidate theory would actually prosper. Even so, it may be said that certain philosophical concepts could be used

[22] *Ibid*., 25.
[23] *Ibid*., 32.

to inform any theory regarding that nature of AMAs. Among these include the standard views on the nature of agency, rational agency, moral agency, and so on.

Philosophical Ideas re the Nature of Agency

The question about agency has a long history in philosophy. Philosophers from the ancient times, for instance, have debated about its nature and function. Some have claimed that a necessary requirement for agency is having a sort of mental state such as intentionality and awareness. Others have added the condition of voluntariness as a primary requisite for it. Still others have argued that rationality should be a requirement for agency. For some, agency coalesces with accountability and responsibility. Meanwhile, others claim that it coalesces with causality. But whatever the conditions for agency one might offer, and in this work we will take the standard view, it is hard to deny the conceptual link it has with the ability to perform certain actions.

According to many philosophers, agency requires that some particular entity, be it a human or nonhuman entity, is capable of performing some action. In this regard, we could say that agency is a two-fold causal relation between an entity and an action.[24] It is a causal relation in that the entity is taken to be the source or the initiator of the action. Let us set aside for now the nature of the entity doing the action, and focus on the nature of the action itself.

When we ordinarily talk about actions, we often oscillate between two separate things. We might think of actions as things that merely happen to us, or else things that we actually do.[25] Philosophers call these two as events and actions, respectively.[26] To explain this distinction, consider the act of breathing and the act of writing.[27]

---

[24] Markus Schlosser, "Agency," Stanford Encyclopedia of Philosophy, ed. Edward Zalta, 2012. http://plato.stanford.edu/archives/fall2015/entries/agency/ (accessed November 1, 2015).

[25] George Wilson and Samuel Shpall, "Action," Stanford Encyclopedia of Philosophy, ed. Edward Zalta, 2012. http://plato.stanford.edu/entries/action (accessed May 24, 2015).

[26] There is a whole philosophical debate regarding the nature of actions, which will be beyond the scope of this paper. For more information about this debate, see Wilson and Shpall, "Action."

[27] Kenneth Einar Himma, "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties must an Artificial Agent have to be a Moral Agent?," *Ethics and Information Technology* 11 (2009): 19-29. https://www3.nd.edu/~dimmerma/teaching/20402-

Writing could be considered as an action as its performance is causally dependent on the person who is doing the writing. But now we have a problem, we could also say that breathing is an action, for the same reason—it is causally dependent on the entity that does the breathing. But the difference between the two "actions" is that one requires a corresponding mental act that causes the entity to do the action, while the other does not necessarily require it. Breathing just happens—it is an event. Writing, on the other hand, is more intentional in that it is more of an agent's action.

The theory that actions have a corresponding mental act, which causes an agent to do an action, has been supported by many philosophers.[28] This theory tells us that an action requires a mental phenomenon to go along with it, acting as a causal nexus. Any "action" which does not have this mental component does not qualify as an action.[29] But we could be neutral as to the specific kind of mental state that actions should be dependent on, since these could either be a willing, volitional, or a belief-desire pair of mental states. The important thing is that these are of the intentional kind (i.e., they should be directed to, or be about, something else).[30] Furthermore, we could also stay neutral about the very nature of mental phenomenon. We do not need to delve into the question of whether it is physically explainable or not.[31]

But how does this conception of actions relate to agency? We could take this idea as an answer:

> Agency, as a conceptual matter, is simply the capacity to cause actions—and this requires the capacity to instantiate certain intentional mental states... The most common view... is that it is a necessary condition of

03/locked%20PDFs/Himma09_ArtificialAgencyConsciousnessAndTheCriteriaForMoralAgency.pdf (accessed May 7, 2015).

[28] This theory was famously defended by Donald Davidson, *Essays on Actions and Events* (Oxford: Oxford University Press, 1980).

[29] A long standing debate in the philosophy of mind is whether these mental phenomena are, in principle, physically grounded. This work would be neutral about this issue. For another take on the issue, see Napoleon Mabaquiao, Jr., *Mind, Science and Computation* (Manila: Vibal Publishing House, Inc., 2012).

[30] Compare with Kenneth Einar Himma, "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties must an Artificial Agent have to be a Moral Agent?"

[31] Again, a good guide for this issue is Napoleon Mabaquiao, Jr., *Mind, Science and Computation*.

> agency that the relevant *mental* states are capable of causing performances... Thus, the following constitutes a rough but accurate characterization of the standard view of agency: X is an agent if and only if X can instantiate intentional mental states capable of directly causing a performance.[32]

One further idea that we could draw from this characterization is how agency relates to rational agency. From the definition above, we get the notion that some entity is an agent so long as it performs an action brought about by some intentional state. But it does not follow from this that all beings who have intentional states automatically qualify as rational agents. For example, dogs are agents in that they can perform certain actions, which are brought about by certain intentional states. However, unlike humans, they are said to be not rational agents.

A fundamental difference between human beings and dogs has something to do with rationality (i.e., the ability to deliberate the reasons behind an action). While dogs can have intentional actions, they could not really deliberate on their reasons for doing some action. In contrast, humans could actually deliberate the reasons behind their actions. Thus, this process where an agent deliberates the reasons for an action could be considered as an integral condition for something to count as a rational agent.[33] This conception of rational agents would therefore amount to this: something is a rational agent if and only if for some action that this agent performs, the agent has deliberated on the reasons for doing such an action. Furthermore, this conception parallels another view about rational agents.

Some theorists view rational agency not just as an ability to deliberate on the reasons for some action, but also as the ability to modify, shape, and control the environment that agents are situated in.[34] In this view, not only do agents have reasoning and deliberating processes, they also have preferences over certain possible outcomes. Thus, a rational agent could be understood as a being that acts in its own best interest. This means that for something to even count as a rational

---

[32] Kenneth Einar Himma, "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties must an Artificial Agent have to be a Moral Agent?," 20-21.

[33] *Ibid*., 20.

[34] Wiebe van der Hoek and Michael Woolridge, "Towards a Logic of Rational Agency," *Logic Journal of IGPL* 11, no. 2 (2003): 133-157. http://www.cs.ox.ac.uk/people/michael.wooldridge/pubs/igpl2003a.pdf (accessed May 21, 2015).

agent, it should be capable of calculating and choosing its own actions so that the outcomes would be optimized with respect to its preferences.

With regards to the question about the nature of the agent itself, some theorists have distinguished between two types of agents: natural agents, on the one hand, and artificial agents, on the other hand.[35]

Natural agents are those whose existence is accounted for in biological terms. These agents are considered as part of some biological specie, and are said to be mostly products of biological reproductive capacities. It is quite obvious that humans and animals fall under this category. Meanwhile, artificial agents are non-biological entities that satisfy the criteria of agency. These beings are "manufactured by intentional agents out of pre-existing materials external to the manufacturers."[36] These agents are also called "artifacts," since they are artificially manufactured. But though they are artificially produced, these artifacts might still be capable of performing intentional actions. Furthermore, it may be argued that these artificial agents could also deliberate the reasons for their actions—hence, can be considered as rational agents as well. Among those that may be included in this category are sophisticated computers, intelligent systems, and robots that are able to perform actions caused by intentional mental states. Now that we have a good handle of the concept of agency, let us turn our focus on the concept of moral agency.

According to some philosophers, moral agents are those entities whose actions and behaviors are subject to moral requirements. This means that, under certain ethical standards, moral praise or blame could be ascribed to the actions of these agents. But what are the conditions for someone or something to be considered as a moral agent? Again, this might serve as an answer:

> The conditions for moral agency can thus be summarized as follows: for all X, X is a moral agent if and only if X is… an agent having the capacities for…

---

[35] Relate this with Kenneth Einar Himma, "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties must an Artificial Agent have to be a Moral Agent?" Note that these categories are not exclusive and exhaustive. For instance, clones exhibit both the properties of being natural and artificial in a certain sense. Also, an all-perfect being, like the Judeo-Christian notion of a God, is said to be one of those agents that does not clearly fall under the said dichotomy.

[36] *Ibid*., 21.

> making free choices... deliberating about what one ought to do, and... understanding and applying moral rules correctly in paradigm cases.[37]

One crucial thing to note in this definition is the idea that moral agents are capable of deliberating about what one ought to do. This idea refers to the capacity of agents—rational agents at that—to conduct moral reasoning.

Moral reasoning is typically understood as the process of reasoning by which behaviors and actions are adjudicated or justified for their ethical worth (i.e., on whether these things are in accordance or in violation of certain ethical standards and practices).[38] In general, this reasoning process is comprised of three components, namely: the ethical standards that act as the basis for moral judgments, the relevant facts of a specific context that is under consideration, and the ethical judgment to be derived from these two. Velasquez explains this as follows:

> Moral reasoning always involves three components: (1) an understanding of our moral standards and what they require, prohibit, value, or condemn; (2) evidence or information about whether a particular person, policy, institution, or behavior has the features that these moral standards require, prohibit, value, or condemn; and (3) a conclusion or moral judgment that the person, policy, institution, or behavior is prohibited or required, right or wrong, just or unjust, valuable or condemnable, and so on.[39]

Other theorists, Gallagher for example, claim that aside from the ability to conduct moral reasoning, moral agents should also be "capable of being responsible for their actions, whether their actions are moral or immoral."[40] To arrive at this definition, Gallagher employs the six conditions of moral personhood, which coincides with the six conditions of personhood proposed by Dennett.[41]

---

[37] *Ibid*., 29.

[38] Manuel G. Velasquez, *Business Ethics: Concepts and Cases*. 7th ed. (New Jersey: Pearson Education, Inc., 2012), 45.

[39] *Ibid*., 45-46.

[40] Shaun Gallagher, "Moral Agency, Self-Consciousness, and Practical Wisdom," *Journal of Consciousness Studies* 14, nos. 5-6 (2007): 200. http://www.ummoss.org/gallagher07jcs*.pdf (accessed May 26, 2015).

[41] For a full discussion of his six proposed conditions for (moral) personhood, see Daniel Dennett, "Conditions of Personhood," in *Brainstorms: Philosophical Essays on Mind and*

Borrowing from Dennett's criteria, Gallagher[42] states that an entity must first possess rationality for it to qualify as a moral agent. This requirement correlates with the concept of rational agency discussed earlier. Second, it should be possible to attribute different states of intentions or consciousness, which Dennett calls our ability to take an "intentional stance," to such kinds of entities. The third criterion pertains to the manner by which others treat these entities (i.e., certain attitudes or stances could be adopted by others towards it).[43] So, for an agent to be considered of the moral kind (e.g., persons), "we have to [first] treat it as a person... with respect or, as the case may be, hostility."[44] Another condition for moral agency is the ability to reciprocate the same attitudes or stances. This means that such types of beings should be able to return back and adopt the stances identified by Dennett for his third condition to other supposed moral agents. Meanwhile, the fifth criterion refers to the ability of an agent to (verbally) communicate with others (i.e., moral agents ought to have some linguistic capabilities).[45]

Gallagher notes that the second, third, fourth and fifth conditions relate to the social dimensions of an agent, since these might be seen as prerequisites to any interpersonal relations. On the other hand, putting in the first condition in the set of conditions necessitate a being to be self-conscious[46] for it to be considered a moral

---

Psychology (Cambridge: MIT Press, 1978), 175-196. http://philpapers.org/archive/DENCOP.pdf (accessed May 26, 2015).

[42] Shaun Gallagher, "Moral Agency, Self-Consciousness, and Practical Wisdom," 200.

[43] In relation to this, Dennett further clarifies that "it is not the case that once we have established the objective fact that something is a person we treat him or her or it a certain way, but that our treating him or her or it in this certain way is somehow and to some extent constitutive of its being a person." See Daniel Dennett, "Conditions of Personhood," 177-178.

[44] Shaun Gallagher, "Moral Agency, Self-Consciousness, and Practical Wisdom," 200.

[45] The fifth condition is highly, and was criticized by some philosophers. Some have argued that linguistic capacity is derivative, and hence not fundamental, to rationality. For example, see David Hugh Mellor, *Matters of Metaphysics* (Cambridge: Cambridge University Press, 1991), 30-60. In the said text, Mellor suggests that language is grounded on the capacity for agent to have beliefs. Having beliefs, therefore, is necessary to have linguistic capacities. For this paper, we could remain neutral about this issue.

[46] Note that self-consciousness maybe considered a type of consciousness. For instance, see Robert James M. Boyles, "Artificial Qualia, Intentional Systems and Machine Consciousness," in *Proceedings of the DLSU Research Congress 2012*, 110a-110c. (Manila: De La Salle University-Manila, 2012). https://philpapers.org/archive/BOYAQI.pdf (accessed May 26, 2015). In a way, this could be related to the goal of modeling AMAs as there is a distinct research field, machine

agent. This is considered as the sixth condition. Note that self-consciousness is defined as higher-order reflective mental processes.[47]

It might be argued that Gallagher's six conditions cash out the previously discussed idea that the behaviors and actions of moral agents must be subjected to moral scrutiny. This means that the specified conditions must first obtain for someone, or even something, to be considered a moral agent.

We might notice that the notion of moral agency is conceptually connected to the idea of responsibility and accountability. Moral standards govern the actions of these agents; this implies that, as moral agents, agents have ethical obligations and duties. A moral agent deserves blame, if not punishment, anytime that it violates an ethical obligation. Conversely, it warrants praise every time that it "sacrifice[s] important interests of her own in order to produce a great moral good that... [it] was not required to produce."[48] It could be said that the evaluation of the actions of moral agents (i.e., whether they warrant praise or blame) is defined by moral standards. Such standards dictate which actions are morally acceptable and which are not, and this also entails that moral agents are morally accountable for their behavior.[49]

In terms of developing a theory that accounts for the nature of artificial moral agency, the contributions of philosophers, like the ones mentioned above, could somehow serve as the initial building blocks. For instance, the idea of ascribing moral praise and blame to an AMA is somehow grounded on the basis of it being a rational, artificial, and moral agent (i.e., with the capability of performing actions). Furthermore, the goal of fostering machines that follow a set of ideal ethical

---

consciousness, which focuses on the development of sophisticated machines that possess artificial qualia.

[47] This characterization of self-consciousness reminds us of what Frankfurt calls second-order volitions. See Harry G. Frankfurt, "Freedom of the Will and the Concept of a Person," *The Journal of Philosophy* 68, no. 1 (1971): 5-20. http://verybadwizards.com/s/Frankfurt.pdf (accessed June 2, 2015).

[48] Kenneth Einar Himma, "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties must an Artificial Agent have to be a Moral Agent?," 22. Some philosophers have noted, however, that the mere performance of one's own (expected) duties does not merit any moral evaluation.

[49] It is a common practice to use moral accountability and moral responsibility interchangeably. See Andrew Eshleman, "Moral Responsibility." Stanford Encyclopedia of Philosophy, ed. Edward Zalta, 2014. http://plato.stanford.edu/entries/moral-responsibility (accessed June 2, 2015).

principles is not that farfetched given that moral agents, regardless if such is of the artificial kind, would have to be morally responsible for its actions. However, it might be the case that the said findings of philosophers would not all be applicable in explaining the nature of AMAs. The task, then, is to build on top of such philosophical foundations, if not demolish and laydown new ones.

Conclusion

The ideas enumerated earlier are only some of the many contributions of philosophers that could provide invaluable guidance to those who are doing research on the nature of artificial moral agency. The different ways of accounting for agency, rational agency, and moral agency, among others, maybe factored in by machine ethicists in their search for a final theory on the nature of AMAs. At the very least, the said philosophical insights may be treated as signposts for further research on how to truly account for the latter.

Works Cited

Ackerman, Evan. "Robot Takes on Landmine Detection While Humans Stay Very Very Far Away," *IEEE Spectrum*, January 23, 2014. http://spectrum.ieee.org/automaton/robotics/military-robots/husky-robot-takes-on-landmine-detection-while-humans-stay-very-very-far-away (accessed June 25, 2015).

Anderson, Michael, and Susan Leigh Anderson. "Machine Ethics: Creating an Ethical Intelligent Agent," *AI Magazine* 28, no. 4 (2007): 15-26. http://www.aaai.org/ojs/index.php/aimagazine/article/download/2065/2052 (accessed July 1, 2015).

Bostrom, Nick, and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William Ramsey. New York: Cambridge University Press, 2013. http://intelligence.org/files/EthicsofAI.pdf (accessed July 28, 2014).

Boyles, Robert James M. "Artificial Qualia, Intentional Systems and Machine Consciousness." In *Proceedings of the DLSU Research Congress 2012*, 110a-110c. Manila: De La Salle University-Manila, 2012. https://philpapers.org/archive/BOYAQI.pdf (accessed May 26, 2015).

Davidson, Donald. *Essays on Actions and Events*. Oxford: Oxford University Press, 1980.

Dennett, Daniel. "Conditions of Personhood." In *Brainstorms: Philosophical Essays on Mind and Psychology*, 175-196. Cambridge: MIT Press, 1978. http://philpapers.org/archive/DENCOP.pdf (accessed May 26, 2015).

Eshleman, Andrew. "Moral Responsibility." Stanford Encyclopedia of Philosophy, edited by Edward Zalta, 2014. http://plato.stanford.edu/entries/moral-responsibility (accessed June 2, 2015).

Frankfurt, Harry G. "Freedom of the Will and the Concept of a Person," *The Journal of Philosophy* 68, no. 1 (1971): 5-20. http://verybadwizards.com/s/Frankfurt.pdf (accessed June 2, 2015).

Franklin, Stan 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.

Friedman, Lauren. "IBM's Watson May Soon Be The Best Doctor In The World." *Business Insider*, April 22, 2014. http://www.businessinsider.com/ibms-watson-may-soon-be-the-best-doctor-in-the-world-2014-4 (accessed June 25, 2015).

Gallagher, Shaun. "Moral Agency, Self-Consciousness, and Practical Wisdom," *Journal of Consciousness Studies* 14, nos. 5-6 (2007): 199-223. http://www.ummoss.org/gallagher07jcs*.pdf (accessed May 26, 2015).

Harding, Luke. "At least 23 Die as Driverless Train Crashes into Maintenance Truck." *The Guardian*, September 23, 2006. http://www.theguardian.com/world/2006/sep/23/germany.topstories3 (accessed May 7, 2015).

Himma, Kenneth Einar. "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties must an Artificial Agent have to be a Moral Agent?," *Ethics and*

Information Technology 11 (2009): 19-29. http://www3.nd.edu/~dimmerma/teaching/20402-03/locked%20PDFs/Himma09_ArtificialAgencyConsciousnessAndTheCriteriaForMoralAgency.pdf (accessed May 7, 2015).

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

Mellor, David Hugh. *Matters of Metaphysics*. Cambridge: Cambridge University Press, 1991.

Moor, James H. "The Nature, Importance, and Difficulty of Machine Ethics." In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 13-20. New York: Cambridge University Press, 2011.

Noorman, Merel. "Computing and Moral Responsibility." Stanford Encyclopedia of Philosophy, edited by Edward Zalta, 2012. http://plato.stanford.edu/entries/computing-responsibility (accessed June 2, 2015).

Olsen, Jan Kyrre Berg, Stig Andur Pedersen, and Vincent F. Hendricks, ed. *Blackwell Companions to Philosophy: A Companion to the Philosophy of Technology*. West Sussex: Blackwell Publishing Ltd., 2009.

Schlosser, Markus. "Agency." Stanford Encyclopedia of Philosophy, edited by Edward Zalta, 2012. http://plato.stanford.edu/archives/fall2015/entries/agency/ (accessed November 1, 2015).

Shulman, Carl, Henrik Jonsson, and Nick Tarleton. "Machine Ethics and Superintelligence." In *Proceedings of the AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference*, edited by Carson Reynolds and Alvaro Cassinelli. Tokyo: AP-CAP, 2009. http://ia-cap.org/apcap09/proceedings.pdf (accessed June 2, 2014)

Sullins, John P. "When Is a Robot a Moral Agent? International Review of Information," *Ethics* 6 (2006): 23-30. http://www.i-r-i-e.net/inhalt/006/006_Sullins.pdf (accessed June 2, 2014).

_____. "Artificial Moral Agency in Technoethics." In *Handbook of Research on Technoethics*, edited by Rocci Luppicini and Rebecca Adell, 205-221. Hershey: IGI Global Information Science, 2009.

van der Hoek, Wiebe, and Michael Woolridge. "Towards a Logic of Rational Agency," *Logic Journal of IGPL* 11, no. 2 (2003): 133-157. http://www.cs.ox.ac.uk/people/michael.wooldridge/pubs/igpl2003a.pdf (accessed May 21, 2015).

Velasquez, Manuel G. *Business Ethics: Concepts and Cases*. 7th ed. New Jersey: Pearson Education, Inc., 2012.

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

Wilson, George, and Samuel Shpall. "Action." Stanford Encyclopedia of Philosophy, edited by Edward Zalta, 2012. http://plato.stanford.edu/entries/action (accessed May 24, 2015).

Yudkowsky, Eliezer. "Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures." Machine Intelligence Research Institute, 2001. http://intelligence.org/files/CFAI.pdf (accessed June 2, 2014).