Moral Artifacts: Enframing, "Ready-Ethics," and Anthropocentrism

Jun Kyung You

Supervisor: Prof. Mueller

"Fall/Winter 2018-19"

Word Count: 16080

1. Introduction

It is probably unwise to criticize an author exclusively based on the contents of what she wrote in the abstract and introduction of her paper. But there are very troubling aspects to be found in some of the abstracts and introductions of recent Ethical AI research.

> "The future will see autonomous machines acting in the same environment as humans … thus hybrid collective decision-making systems will be in great need."[1]
> "We propose to replace preference aggregation with an appropriately developed value/ethics/preference *fusion*, an operation designed to ensure that agents' preferences are consistent with their moral values and do not override ethical principles."[2]
> "Arguably the main obstacle to automating ethical decisions is the lack of a formal specification of ground-truth *ethical* principle, which have been the subject of debate for centuries among ethicists and moral philosophers … Dwork et al. concede that "when ground-truth ethical principles are not available, we must use an 'approximation as agreed upon by society.'""[3]

Here, these words reek of the ethical baggage being smuggled through. Upon reading them, you, a philosopher, perhaps can also visualize the writers standing in front of a customs officer and presenting their papers to him, waiting for the approval stamp; praying that he jumps across these words, accepting them as mere "formalities," so that he only asks the questions they can confidently answer. But these words carry too much weight for us to simply treat them as "formalities." They reveal certain attitudes of these researchers regarding 1. the nature of technological development; 2. how ethical decision-making works for humans (or how the decision-making should look like in artificial agents); and 3. their attitude towards Ethics in general. We cannot help but question these attitudes.

* * * *

---

[1] Greene, J. et al, "Embedding Ethical Principles in Collective Decision Support Systems," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016. p.4147.
[2] Ibid.
[3] Noothigattu, R. et al., "A Voting-Based System for Ethical Decision Making," *Proceedings of Autonomous Agents and Artificial Intelligence (AAAI) Conference*, 2018. p.1.

This paper starts by identifying and criticizing these attitudes, which can be attributed to the vast majority of Ethical AI researchers. The first attitude concerning the nature of technological development is examined by using the idea of Heideggerian "Enframing;" the second attitude concerning ethical decision-making is referred to as the "Subsumptivist Generalist Position" or SGP; the third attitude on Ethics in general is referred to as the "Instrumentalist view of Ethics." I explain each of these, show how they manifest in Ethical AI research, and proceed to criticize each of these in Section 2. In subsequent sections, I consider my criticism's implication to the field of Ethical AI. In Section 3, I consider the Moral Particularist approach to designing ethical artificial agents, and how this might be a more suitable approach to machine ethics compared to that which is driven by the attitudes described in Section 2. A rare case in which this Particularist implementation is experimented is also considered, to vindicate the practical possibility of the Particularist approach. In Section 4, I consider the full implications of an ethical artificial agent technology. I reject that a supposed ethical artificial agent can have moral agency that would make it acceptable to replace human moral agency, even if it is supposed that the artificial agent has full moral competence.

2. "Technological Over the Ethical"—Survey of the Three Attitudes

2.1. Enframing

Consider this passage, along with the first of the quotes in the introduction, which launches a celebrated Ethical AI research article featured in *Nature* few months ago:

> "With the rapid development of artificial intelligence have come concerns about how machines will make moral decisions, and the major challenge of quantifying societal expectations about the ethical principles that should guide machine behavior"[4]

Here we are presented with two problems that the researchers intend to solve via this experiment. It is the task and in the interest of the scientist to figure out *how* these problem should be solved, whereas it is the task and in the interest of the philosopher to figure out,

[4] Awad, E. et al. "The Moral Machine Experiment." *Nature* vol. 563, 2018. p.59.

among other things, *who* is presenting these problems to the scientists, and *why* this problem became so important to all of us and not just the scientists. Who is driving the "rapid development of artificial intelligence" that this has become an urgent matter for the scientists to address? Consider the mystery of who or what is behind the drive of AI development along with the even more mysterious *certainty* that these researchers have regarding the shape that the future will take once their efforts have come into fruition. Why are they so sure that "the future will see autonomous machines acting in the same environment as humans"? From where do we get this certainty, I dare say righteousness, regarding technology, and the future that it will bring to us? Why do we view technological development as something akin to a natural process, like the strike of thunder that we have no control over, rather than as we view any other human projects which, in fact, we do have control over? These questions hint at a certain attitude of these scientists concerning the nature of technological development, an attitude which I find repulsive.

This *attitude* is more complicated than simply believing that our life is made better through technological development. There is a deeper and radical metaphysical and normative claim to this *attitude*, which is that technology is the only way to approach the truth, and that this is the best way to improve our condition. The result of this *attitude* is that devoting oneself to technological development is the most valuable vocation, that it is wrong to obstruct such progress, and that technological progress *will occur* and bring the benefits as we imagine them. Exactly how and why technology achieves and asserts this kind of dominance is explained by Heidegger in his *The Question Concerning Technology*. One of the key ideas in this essay, enframing, is perhaps the best expression of the *attitude* I hint to. So, I engage in an exposition of his critique of technology and finally arrive at his idea of enframing so that we can begin fleshing out this problematic *attitude*.

2.1.1. Exposition of Heideggerian *Enframing*

Heidegger's project in this essay is to conduct a "*questioning* concerning technology."[5] Heidegger proposes that in order to conduct such a questioning, we must open our existence to a "free relationship" with technology, meaning that we should clearly

[5] Heidegger, M. "The Question Concerning Technology." *The Question Concerning Technology and Other Essays*. Garland Publishing, 1977. p.3.

perceive the whole significance of technology first.[6] To do so, he begins with an intuitive definition of technology. There are two parts to this definition: 1. "Technology is a means to an end"; 2. "Technology is a human activity." This pair of definitions is called the "instrumental and anthropological definition of technology."[7] There are no doubts to the correctness of these definition, as Heidegger concedes, but he does not take the correctness of this definition to suffice as the essence of technology, and suggests that we should consider what it means for technology to be instrumental.[8] Heidegger notes that "wherever instrumentality reigns, there reigns causality," as there is always an end that is caused by the instrumental.[9] In other words, an instrument is only used *for* another thing, and the design of the instrument is meant to *cause* to this specific end. So, causality is invariably shared by all instruments. Therefore, he brings us to consider causality. After a discussion of the four kinds of causes laid out by Aristotle (the specifics of these four causes are not important here), Heidegger claims that what all these four different kinds of causality have in common is their act of "bringing-forth," or "bringing something into appearance," not only in the sense of creating something via handicraft, but also of the "growing things of nature."[10] But this act of bringing-forth can only make sense insofar as there is something to be brought into existence out of non-existence. As creating existence out of utter non-existence is impossible, it must be the case that the supposed "non-existence" was in fact concealing a realizable possibility. So, we can conclude, Heidegger argues, that bringing-forth is the act of "revealing" something that was concealed so far.[11] The careful selection of words by Heidegger is directed towards a specific sense: when we use technology to create things, we do not do so out of nothing. With technology we introduce our viewpoint about what the world looks like, and through shaping the world in a way that fits this specific kind of viewpoint, we assume and vindicate our viewpoint at the same time. *Technology is both the postulation and proof of certain kinds of truths.*

So, technology, insofar as it is understood as "the possibility of all productive manufacturing," is a revealing[12]—it is an industry of the truth. It is not only the means by

---

[6] Ibid.
[7] Heidegger, p.4-5.
[8] Heidegger, p.6.
[9] Ibid.
[10] Heidegger, p.10-11.
[11] Ibid.
[12] Heidegger, p.12.

which we fulfill our needs, and it is not only the art of producing tools of sustenance. Through engaging in such productive manufacture, technology establishes truths and therefore *reveals* what in fact *is*. This, indeed, sounds like a strange portrayal of technology, especially for the modern one, as we usually think that "modern technology is something incomparably different from all earlier technologies because it is based on modern physics as an exact science."[13] I think, for those who also feel this strangeness, it is believed that science itself does the revealing, while technology merely uses the truths uncovered by science. This thought supposes that science, or the human will to truth itself, is the only driving factor that has control over the direction of technology, as technology is merely the means that utilize the truths uncovered by science for the purpose of our comfort and survival. Technology as an instrument is designed with the goals directed by science, these would say. The genius of Heidegger is to see that there is another direction of causality in the development of technology, that is pushed forward not by our will to truth but by the requirement of modern technology as an *instrument*—modern technology is an autonomous vehicle, if you will. This peculiar trait of modern technology, which is that it has an additional direction of causality, makes modern technology "alone that is the disturbing thing, that moves us to ask the question concerning technology per se."[14]

From here, Heidegger moves on from what detractors might consider as Greek wordplay, to directly engage in a phenomenological investigation of modern technology: the "revealing that rules in modern technology is a challenging [*Herausfordern*], which puts to nature the unreasonable demand that it supply energy that can be extracted and stored as such."[15] "Agriculture is now the mechanized food industry. Air is now set upon to yield nitrogen, the earth to yield ore, ore to yield uranium," etc.[16] Everything is seen as a "standing-reserve,"[17] or *something that is organized beforehand so that it can be readily used for a further purpose*. Perhaps Heidegger's discussion of the Rhine's transition into a "standing-reserve" is the clearest illustration he gives of this:

"The hydroelectric plant is set into the current of the Rhine. It sets the Rhine to

---

[13] Heidegger, p.14.
[14] Ibid.
[15] Ibid.
[16] Heidegger, p.15.
[17] Heidegger, p.17.

> supplying its hydraulic pressure which then sets the turbines turning. This turning sets those machines in motion whose thrust sets going the electric current for which the long-distance power station and its network of cables are set up to dispatch electricity. In the context of the interlocking processes pertaining to the orderly disposition of electrical energy, even the Rhine itself appears as something at our command. The hydroelectric plant is not built into the Rhine River as was the old wooden bridge that joined bank with bank for hundreds of years. Rather the river is dammed up into the power plant. What the river is now, namely, a water power supplier, derives from of the essence of the power station … let us ponder for a moment the contrast that speaks out of the two titles, "The Rhine" as dammed up into the *power works,* and "The Rhine" as uttered out of the *art* work … But, it will be replied, the Rhine is still a river in the landscape is it not? Perhaps. But how? In no other way than as an object on call for inspection by a tour group ordered there by the vacation industry."[18]

Heidegger notices, in the rise of optimizing, stockpiling, and expediting that is characteristic of modern technology, a shift in viewpoint occurs, which ceases to see nature as the source of truth or what dictates our way of life, but rather as an exploitable system that should be organized and made ready for usage towards a further purpose—including the people who populate that world (hence the term "human resource"). Consider the difference between the wooden bridge over the Rhine and the hydroelectric plant in the Rhine and this point may be clearer. The wooden bridge was used for people to overcome the bridge, because the Rhine was a fact of life for the people who lived around it, and technology (the bridge) was used for people to be able to live *nevertheless* the existence of the Rhine. The Rhine *itself* is taken for granted, and no efforts are made to alter the essence of the Rhine. The Rhine was part of the landscape in the truest sense of the term. The dam connected to the power plant, on the other hand, was built with the intention to exploit the hydroelectric potential in the Rhine. Further, it alters the environment of the Rhine in a very specific manner so that the Rhine itself can be optimized in producing this hydroelectric resource. The essence of the Rhine is only interpreted from this viewpoint of *resources*, and as something that is required for an additional end. Even the Rhine being part of the landscape is not left unexploited, as it is set

[18] Heidegger, p.16.

on display for use by the "vacation industry." This purposeful organization of the world in readiness of a further end is the character of modern technology. When this organization is applied to nature, the relationship that technology holds with nature is reversed. Nature is no longer the source of the immutable truth that restricts the means that technology should take. The requirements of technology displace nature as the officeholder of the immutable truths and restrict the form that nature can take. Technology, as it assumes and vindicates a specific kind of viewpoint of the world via its completion, also makes demands that nature take a specific kind of posture—as a "standing-reserve." Science, in the face of this dominating nature of technology, also changes its purpose from it being a tool for interpreting *what is* of the world, to *how* this world could be organized into given the direction that technology imposes. Seeing a further purpose for everything, and that everything should be organized to maximize the fulfillment of that further purpose, is the character of modern technology according to Heidegger. This character of modern technology which causes the reversal, the tendency of technology to organize the world for a further end, is named *enframing*.[19] The human affirmation of technology's development as it is directed by technology's requirements can therefore also be understood as a manifestation of enframing.

2.1.2. *Enframing* in Machine Ethics—"Ready-Ethics"

As we now understand the road that Heidegger takes to show what enframing is, we are naturally brought to consider how this idea could be useful in interpreting Ethical AI research. One major trouble in bringing this concept to use for our purposes is that enframing as Heidegger originally formulates it describes and criticizes the posture of modern technology in general. It is not in my interest, nor in alignment with my stance, to present a critique against modern technology in general via fully imbibing in Heidegger's suspicion against technology. But I do think this concept is powerful and useful in understanding the milieu from which problematic attitudes in certain Ethical AI research are produced. To make enframing applicable, therefore, I take another step to clarify the sense I wish to utilize in Heidegger's enframing.

Consider this (comparatively) straightforward rendition of enframing: "Enframing might be thought of as the ordaining of destining that establishes the technological clearing as

[19] Heidegger, p.19.

the one dominant picture, to the exclusion of all others."[20] The essence of enframing is the technological structure's dominant and aggressive posture in the matter of "clearing," which is "revealing," of the metaphysical and normative. In other words, enframing is a structure's assertion of its specific means of interpreting and creating the world as the correct and right ways to do so. Heidegger introduces enframing with the intention to say that this kind of posture is firmly engrained in modern technology, as it relies on modern physics as an "exact science." The posture of modern technology as enframing is not tied to technology itself, but only to "modern" technology. Modern technology is an extension of modern physics, which purports to be able to provide the most truthful explanation of the nature of the world (and which we *in fact* accept as the authority of physical facts). So, it seems apparent that modern technology, which superficially serves modern science, should also assert dominant authority in determining what should be done in order to realize the facts presented by modern science. Modern technology, according to this account, seems destined to dominate.

This scent of technological determinism prompts me to depart from Heidegger in the context of machine ethics. I argue it is possible to have machine ethics (which is, of course, an apparatus of modern technology) that is not guided by this attitude of enframing, and it is not necessarily through the power of the aesthetic that grows in its "saving power" as the technological permeates all aspects of our lives.[21] I find the possibility of "saving power" in the disenfranchisement of enframing itself. Recognition of a structure that operates beyond our consciousness gives us also the ability to counteract against that structure. Heidegger provides us the awareness that technology is a structure that governs the direction of our pursuits towards a certain end. The recognition, which is also a revealing, provides us with "saving power," which grows in strength not because of an alternative structure, which for Heidegger is aesthetics, but through the expansion of the reach of human consciousness over this structure. Then mastery of this structure into mundaneness follows: 1. After recognition, we figure out just how far we went to let this structure govern our pursuits; 2. We rescind the pursuits that we realize were misguided based on our values; 3. The structure is then demoted and tagged as ideology—an "interpretation" (the most dramatic case of this in the West is probably Christianity). I believe enframing that has hitherto influenced machine ethics can

---

[20] Wheeler, M., "Martin Heidegger," *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), Edward N. Zalta (ed.), URL <https://plato.stanford.edu/archives/win2018/entries/heidegger/>. 3.3 Technology. Par. 8.
[21] Heidegger, p.32-35

also be “mastered” this way.

It follows, then, that I will be “mastering” enframing in machine ethics by first recognizing its influence on Ethical AI research. What is the attitude of enframing as it is displayed by Ethical AI researchers? It is to place the technological above and prior to the ethical. Specifically, it is to place the priority of expanding the application of artificial intelligence capability above the comprehensive treatment of relevant ethical questions, or to ask these questions only in the context of expanding AI application. Considering the inception of this field, the incentives for this attitude is clear. AI is a tool meant to replace human agency where it seems valuable to do so. For replacement to be as widespread as possible, it follows that it must be able to replicate what is essential to human agency where needed. Morality is an important aspect of human decision-making; therefore, an artificial agent, in performing its task, should be able to understand our “language” of morality, as it were. In the face of this need, Ethics is expected to be answerable to the demands of technology as a “standing-reserve” just like any aspect of nature. It is expected that, as in one of the quotes above, there is a “formal specification of ground-truth *ethical* principle” ready for the AI researcher, and the history of Moral Philosophy is retrospectively, in light of technology’s demand, interpreted as a process to arrive at this “formal specification,” as if the great names in Moral Philosophy all were in fact looking for this “formal specification” to the good and the bad. There is a hierarchical reversal that perfectly mirrors what happened between Nature and Technology in Heidegger’s Rhine example. Technology is given priority over the ethical. This results in Ethics that is very different from the Ethics that Moral Philosophy pursues. This way of inquiry demands and results in an Ethics as a “standing-reserve.” It is crucial to diagnose this kind of Ethics when we see an instance of it, so let us call this “*Ready-Ethics*.”

A question that can be raised against Heidegger’s Rhine example can also be applied here: “How can one be sure that Nature and Technology, when each takes the lead, ends up producing different outcomes?” It could be that Nature-guided-technology and Technology-guided-nature end up taking the same shape. Perhaps the same could be said for Ethics and “Ready-Ethics.” The former is of no interest to us, so I turn to only the latter. We must examine the practice of “Ready-Ethics” for us to be sure that it is indeed different from Ethics. Let’s look for hints in some defining examples in machine ethics. Allen and Wallach, in their seminal work on machine ethics, give a useful illustration of the technological need to

have artificial agents act under our moral approval: "[In the case of a service robot in a home] Wouldn't the robot need to discern whether an obstacle in its pathway is a child, a pet, or something like an empty paper bag and select an action on the basis of its evaluation?"[22] Yes, indeed. But we also need to ask *why* we affirm this need. Understanding the larger goal of machine ethics helps. The aim of machine ethics, according to Powers, is to "produce an autonomous machine that will behave in ways that humans will find generally acceptable from a moral point of view."[23] This is the answer to the *why*. It is because we need to find the moral valuations of artificial agents acceptable for us to be at ease with them replacing human agency. Here is where the difference of concern in Ethics and "Ready-Ethics" is most clearly revealed. It can be said that Ethics, among other things, is a philosophical investigation to the question that we ask every day regarding our actions: "what should I do here? Is what I have done a good thing?" In the sense that the "should" "remains indeterminate,"[24] and we ask questions regarding the nature of this "should." Further, there is a unified *normative* motivation in asking these questions, which we all share, and therefore is the characteristic of human life. I call this question the *normative question.*[25] "Ready-Ethics," on the other hand, is concerned with finding the best version of artificial agent morality that we can accept. The very first step in "mastering" the enframing in Ethical AI research is to realize that "Ready-Ethics" is driven by a noticeably different inquiry, which is to find instances of normative acceptability, and is driven by an empirical rather than a normative motivation. Despite appearing otherwise, therefore, "Ready-Ethics" is not a branch of Ethics, rather, it is a branch of Artificial Intelligence research where the technological need of the social acceptance of artificial intelligence merely takes on the guise of ethical inquiry.[26]

---

[22] Wallach, W., and Colin A.. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2010. p.15.

[23] Powers, T. M. "Models for Machine Ethics." *Philosophy and Computers*, Vol. 14 No.1, p.4. URL= <https://cdn.ymaws.com/www.apaonline.org/resource/collection/EADE8D52-8D02-4136-9A2A-729368501E43/ComputersV14n1.pdf>

[24] Habermas, J. "On the Pragmatic, the Ethical, and the Moral Employments of Practical Reason." *Justification and Application*. Cambridge, MA: MIT Press, 1993. p.2

[25] One could also say that this is our volition to do good.

[26] Machine ethics is not synonymous to "Ready-Ethics." This is because there can be researches in machine ethics that work at the intersection between these two empirical (how can artificial agents be found morally acceptable) and normative (what is a good thing to do) motivations, namely, by finding ways in which artificial agents can be found acceptable by genuinely attempting at an answer to the normative question. Machine ethics, insofar as it is supposed to be a branch of Ethics, can only be so if the normative inquiries come *prior* to technological needs. It is part of the aim of this paper to argue that all machine ethics research should proceed in this manner as long as we believe that "mastering" over the enframing of Ethical AI research is desirable. I present what I think to be the instances of those kinds of research later in the paper.

Recall that technology as enframing claims dominance in both metaphysical *and normative* matters. Although "Ready-Ethics" is only motivated by the need to find acceptability for artificial agents, insofar as acceptability requires the incorporation of some kind of moral viewpoint into artificial agents, "Ready-Ethics" proposes an answer to normative matters that Ethics is concerned with. Further, "Ready-Ethics," insofar as it is a manifestation of enframing, must assert that the moral viewpoint that it incorporates is a correct one. There is a unified, specific answer that "Ready-Ethics" provides to the normative question as observable from the available Ethical AI research. In the following subsections, I assess the answer to the normative question that "Ready-Ethics" proposes.

2.2 Subsumptivist Generalist Position and Instrumentalism of "Ready-Ethics"

The answer that "Ready-Ethics" provides in face of the normative question cannot be as easily found by looking at articles that propose principles on which AI should be designed[27], etc., and I say this for two reasons. 1. These ethical principles that are provided on behalf of the scientists do not by themselves make the attempt to answer the normative question. These principles take for granted already a certain answer that is provided to the normative question, and *that* answer is what we are interested in. The impressive list of "47 principles"[28] do not include a single account of what kind of an ethical viewpoint the engineer will be taking in pursuing the goods presented in those principles. These principles simply present what is taken to be good as that which should be pursued by the engineers. 2. These principles are intended for the engineers and not the machines. These guidelines of "ethical design" does rule out some specific implementations, but not enough so that the engineer retains freedom regarding the answer to the normative question that will be integrated into the design of the artificial agent. We cannot guess the actual implementation of the "moral machine"—*how* it makes moral decision-making—simply based on the principles that govern the engineers. But we are also concerned with this *how*, since we will be directly interacting with these agents, and will be directly impacted by this *how*, if the situation ever arises.

---

[27] Floridi, L., Cowls, J., Beltrametti, M. et al. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds & Machines* (2018) 28: 689. https://doi.org/10.1007/s11023-018-9482-5

[28] Floridi et al., p.8

In the void of a definitive answer to the normative question, technology as enframing stops consulting Ethics and gives its own answer that is "Ready-Ethics." The requirement of technology that drives itself makes certain options more appealing and accessible to the engineer and researcher. Further, it also begins to doubt the legitimacy of the normative question, and tries to find ways to modify the nature of the question so that it can be answered in the format that will be useful for the advance of technology. But technology cannot testify by itself on what this answer is. This answer, therefore, can be only retrospectively found in Ethical AI research that went far enough to, at least, theorize its instantiation and those that went beyond. There are two aspects of "Ready-Ethics" to be found through the survey of relevant research.

One facet of "Ready-Ethics" can be found in the unified adoption of the Subsumptivist Generalist Position (abbv. SGP), which argues that "the moral status of an act is determined by its falling under a general moral principle."[29] Some readers may recognize this term from the discussion between Moral Generalists and Particularists, and it is from exactly that context I bring this term to characterize an aspect of "Ready-Ethics." I borrow many of the relevant and well-argued propositions made against this stance to show the flaws of SGP after discussing some examples of this in Ethical AI.

Another facet of "Ready-Ethics" is the instrumentalism of the ethical. There are two different but related senses to the word "instrumental" that I capitalize on. The first is that of choosing moral theories that are useful for the advancement of the technological. The means, the ethical, is adjusted depending on the needs of the end, the technological. Why SGP is the dominant viewpoint in "Ready-Ethics" is also clarified. The second is that the ethical is taken to be an instrument that is *readily available* for use by the technological. What this involves is the demotion of the normative question into an empirical question, a question that is readily answerable and therefore its answer readily implementable. Specifically, "Ready-Ethics" makes the mistaken if not naïve assumption that the normative question is answerable based on our conduct in the world.

2.2.1-A The Subsumptivist Generalist Position and its Dominance in AI Ethics

---

[29] Strahovnik, V. "Introduction: Challenging Moral Particularism." *Challenging Moral Particularism*, edited by Mark Norris, Lance et al., Routledge, 2008, p.1.

We should understand what SGP is to find an instantiation of it. Before SGP is explained, we must briefly go over the dispute between the Moral Generalists and Moral Particularists, for the acknowledgement of their dispute is the precondition for truly getting a grasp of SGP. To move on to the matter of artificial agents as soon as we can do so, I have focused on the essence, rather than on the completeness, of the two stances.

The divide between Moral Generalism and Moral Particularism that is relevant for our purpose lies in their disagreement on the role of moral principles in practical moral situations. In other words, they give contrasting answers to the question "How useful is a moral principle in providing justifications for our actions?" My focus on actions here is deliberate. Only a part of the Generalist-Particularist debate is relevant for us, since we are concerned with the practical manifestations of our moral theories in the form of ethical artificial agents. Moral Generalists believe that moral principles are useful in determining what the good thing to do is. The reader may rightfully ask, what is a moral principle? The variety of possible definitions of moral principles corresponds to the level of obscurity of the Generalist-Particularist debate. There are at least six different ideas of moral principles in circulation.[30] As we are dealing with ethical artificial agent design, for convenience's sake, I would like to formulate a simple definition of my own that allows us to access the relevant aspects of the Generalist-Particularist debate so that we can understand SGP. The definition I provide is as follows: moral principles are *generalized*, *exceptionless* imperatives for or against actions (e.g., thou shalt not kill). It would be presumptive to say that all renditions of moral principles must be *exceptionless*,[31] but for the purpose of discussing Ethical AI research, as principles are typically treated that way in that sphere, it is appropriate here to define it as so. In any case, Moral Generalists can be characterized by their unified confidence in the validity of the role of moral theory in moral decision-making[32]—and hence in the role of moral principles formed through such theories.

Moral Particularists believe that moral principles are not required to determine and justify actions, and perhaps that the use of them is fundamentally mistaken. According to Dancy, the "strongest defensible position" of the Moral Particularist position "holds that

---

[30] Schroeder, M. "A Matter of Principle." *Noûs*, vol. 43, no. 3, 2009, p. 569.
[31] Guarini, M. "Particularism and Generalism: How AI can Help us Better Understand Moral Cognition," in *Technical Report for Machine Ethics Symposium, American Association for Artificial Intelligence Fall Symposium*, Nov. 2005. p.1
[32] Dworkin, G. "Theory, Practice, and Moral Reasoning." *The Oxford Handbook of Ethical Theory*, by David Copp, Oxford Univ. Pr., 2011. p. 626

though there may be some moral principles, still the rationality of moral thought and judgment in no way depends on a suitable provision of such things … [and that] Moral principles are at best crutches that a morally sensitive person would not require, and indeed the use of such crutches might even lead us into moral error."[33] In any case, the Particularist suggestion to arrive at the best moral decision, as Dancy puts it, "is to look really closely at the case before one."[34]

While there are significant overlaps between Generalism and Particularism, members of each group are united in their opposition to the strongest version of their counterparts. SGP is the most extreme version of Moral Generalism that Particularists are unanimously opposed to. The unique contribution of Particularists to Ethics is their convincing arguments for reasons to believe that our moral life cannot be fully explained by only moral principles. Due to this contribution, moral theories that completely reject contextualism in applying principles are discredited.[35] It is taken as widely agreed that some qualifications made into moral principles if we are to determine and justify actions through their use. But more details on this contribution, and the implausibility of SGP will come later, as we must first return to Ethical AI for the moment and see why most of the researchers of this field deserve being attributed the stance of SGP.

It is an understatement to say that SGP is an aspect of the Ethical AI discourse—it is what *defines* it now. There seems to be a dominant viewpoint in machine ethics which thinks it is not desirable to approach moral decision-making in any way that is not in the principle-oriented direction.[36][37][38][39][40] This point is perhaps shown most clearly by the fact that

---

[33] Dancy, J., "Moral Particularism", *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/moral-particularism/>.

[34] Dancy, J.. *Moral Reasons*. Blackwell, 1993. p. 63.

[35] Lance, M., Little, M. "From particularism to defeasibility in ethics," *Challenging Moral Particularism*, edited by M., Lance et al., Routledge, 2008. p. 54

[36] Goodall N.J. (2014) Machine Ethics and Automated Vehicles. In: Meyer G., Beiker S. (eds) Road Vehicle Automation. Lecture Notes in Mobility. Springer, Cham. p.7

[37] Anderson, M. et al. "A Value Driven Agent: Instantiation of a Case-Supported Principle-Based Behavior Paradigm." *AAAI Workshops* (2017).

[38] Powers, T.M. "Prospects for a Smithian Machine," International Association for Computing and Philosophy, College Park, Maryland, July 2013.

[39] Powers, T.M. Prospects for a Kantian Machine, IEEE Intelligent Systems 21 (4), 2006.

[40] Note on this list: The footnotes could go on and fill the rest of the sentence, but there is no point in doing so other than in proving my point via visual awe. The other Ethical AI research articles cited so far, as you may have noticed, should also be added to this list. It is perhaps more productive to note of exceptions. As of March 2019, there is only one researcher, Guarini, who tackled the idea of implementing a moral decision-making model under the Particularist moral theory, at least among the machine ethics articles available in English. I discuss more on Guarini's works later in justifying the technical viability of the Particularist approach to ethical artificial agent design.

Asimov's three *laws* of robotics (essentially: 1. A robot should not harm nor allow the harm of harm of humans; 2. A robot should follow human orders unless they contradict 1.; 3. A robot should protect itself unless doing so contradicts 1. or 2.)[41] are not only the fundamental framework of machine ethics in popular culture but also for the professionals in machine ethics. Asimov's laws are provided as a springboard to justify the researcher's motivation in devising moral principles that are sophisticated enough to be compelling, because the three laws, upon some thought, seem obviously insufficient. The direction of inquiry is framed in terms of principles from the very start. The dismissal of Asimov's laws, therefore, are followed by the elaboration of other moral systems that seem more applicable and more impressive. The list of plausible moral systems seems to be, according to Goodall's survey of the moral theories in his exploration of autonomous vehicle decision-making models, "utilitarianism, Kantianism, Smithianism, and deontological" moral theories.[42]

Perhaps the most helpful recapitulation of the ideas presented and implemented in the last ten years in machine ethics is provided by Sharkey (whose comprehensiveness is further vindicated by her choice to begin with Asimov).[43] She makes the distinction between "programming" and "training" approaches in producing ethical artificial agents. In the programming category, which aims to make artificial agents act in morally acceptable ways via programmed moral principles, the most concrete example that she features is presented by Anderson and Anderson, who used inductive logic programming with a machine learning approach so that the artificial agent can extract moral principles from a set of human decisions.[44] The artificial agent can then make choices based on the extracted principles in hypothetical situations. In the "training" category, which sounds very interesting in theory, efforts are made to "raise" artificial agents so that they can be a part of the moral community. The rationale behind this approach is that "programming" an artificial agent to act in a morally acceptable way seems too difficult because moral norms that yield definite judgments are "subtle and context dependent" on the time of application.[45] One notable research that Sharkey cites is that of Riedl and Harrison, who, in their research, told stories that contain sociocultural values to artificial agents, and attempted to instill those values via

---

41 Asimov, I., Runaround. *Astounding Science Fiction 29*, 94-103, Mar., 1942.
42 Goodall, p.7
43 Sharkey, A. Ethics Inf Technol (2017). https://doi.org/10.1007/s10676-017-9425-5
44 Sharkey, p.5
45 Sharkey, p.6

machine learning techniques.[46] The data apparently was preliminary at this time, but for our purposes, being aware of the general direction of this effort is enough.

Regardless of their technical differences, these approaches are both grounded on the assumption that moral decision-making depends on principles or algorithmically interpretable patterns, and that if moral agents are to act acceptably in our moral purview, then they should in one way or another understand those moral principles in their "language," as it were. With the programming approach, this kind of inclination is understandable. A non-principled approach to morality precludes the possibility of morality being directly programmable in terms of laws. The programming approach requires the existence of generalized and exceptionless moral principles so that these moral principles are programmed into the agent when found. It is even argued that attempting to develop artificial agent morality would help illuminate our implicitly held moral beliefs. The research by Anderson et al. is an instance of this case.[47] Here, by applying two different moral principles across situations, an implicit relationship between those two principles is said to have been discovered. Then, the new relationship is taken into account by the artificial agent as an additional principle.

Perhaps the training approach does not require the existence of moral principles in the same sense. But critique of the programming approach such as the one offered by Malle is only based on pragmatic arguments, and so Malle's account by itself does not imply this possibility of non-dependence on exceptionless moral principles. Malle prefers the training approach over the programming approach not because he finds the principle-oriented itself reprehensible, but only because it seems hopelessly difficult to program moral principles that are too, as mentioned above, "subtle and context dependent." Malle's critique seems to be in the right direction, but for the wrong reasons. The reason why the "programming" efforts seem too difficult is precisely because *programming morality requires that morality should be thought only in terms of principles*, when there are aspects and requirements in moral judgments of particular cases that are richer than what can be explained by simply principles. I hypothesize that the primary reason why, despite the lacking nature of moral principles, the principle-oriented approach gained primacy in machine ethics is due to its technical feasibility (more on this point in 2.2.2)—because it is the approach that best *enables*

---

[46] Ibid.

[47] Anderson et al., Anderson S., Armen, C. MedEthEx: A prototype medical ethics advisor. In *Proceedings of the eighteenth conference on innovative applications of artificial intelligence*. Menlo Park, CA: AAAI Press, 2006.

technology to complete itself.

2.2.1-B Deficiencies of Moral Principles in Painting the Moral Life

Regardless of what the actual cause of the predominance of the principle-based approach is, it remains true that this is the predominant approach. I have supplied evidence for thinking so above. Due to its predominance, we can take that "Ready-Ethics" provides the principle-based approach as the only answer to the normative question. So, we can return to examining this answer. To my fortune, many Particularists have done more work than I could ever do on showing the flaws in SGP. Specifically, the flaws of exceptionless moral principles in answering the question of "what to do" has been amply documented in their work. Rather than listing all of them, however, I refer to Dancy's critique of Hare. In this paper he is brought in to show the problems of the *general and exceptionless* aspects of moral principles in justifying moral judgments.

What is presupposed in the belief in the efficacy of moral principles is the belief that a moral imperative drawn from one situation can be rightfully applied to another. This is the *generality* of moral principles. What would this imply? Certainly, this means that both situations shared factors that would contribute to the generation of the same kinds of reasons for finding a certain action acceptable. Therefore, believing in the efficacy of moral principles commits the Generalist to the plausibility of moral principles' applicability in "relevantly similar" situations which are called so due to the situations' similarity in the relevant factors. The task of the Generalist, therefore, includes the presentation of a convincing account of a theory of identifying "relevantly similar" situations. In other words, starting from the particular moral judgments of moral situations, the Particularist asks the Generalist how these judgments can be generalized. The Generalist, then, should explain how this judgment can be generalized to other situations.

How does the effort of the Generalist hold up to this task? Dancy mentions an effort of Hare to define the "relevantly similar" situation: "A situation is relevantly similar to the first if it shared with the first all the properties that were the person's reasons for his original judgment."[48] Dancy immediately proceeds to attack this notion. This notion is too simple to be intuitive, so as to not even require a morally complicated situation to refute it. Let us say I

[48] Dancy, *Moral Reasons*, p.80

eat a candy because I believe it to taste good, and it certainly does once I do eat it. In the next instance, should I eat the candy even after I had too much of it simply because the candy is there and I believe the candy to taste good? Well, since I had too much of the candy, it would be better to not have it. One would want to eat something, regardless of its savoriness, if she is not full. Hare's initial notion is too simple to account for defeating reasons that could be brought up in situations that are different from where the original judgment was made.

Dancy allows for a more favorable interpretation for Hare's suggestion by expanding the "reasons for his original judgment" to include "reasons against," and shows how the definition is still insufficient. Dancy's example involves a man who strikes a woman with his car and puts her in a hospital.[49] We approve the man's decision to make amends by visiting her and paying for special care, and so on. Perhaps the principle that could be extracted from her is that "it is approvable that a person who gets another person involved in a car accident make amends for this other person." But we do not approve the man's doing so, say, with the ultimate ends set at seducing her away from her partner. What should be noted here is that there could be a myriad of defeating reasons introduced into the second situation in place of the intent to seduce the woman which would made the man's decision repulsive. For example, the man may have provided such care only to prevent her from suing him, etc. So, it is unrealistic to say that the lack of the man's intention to seduce the woman was part of the implicit reasons for approving the man's action in the first. In other words, we cannot reasonably say that we approved of the first situation because we *knew* that the man was not intending to seduce the woman, and that this knowledge could have been part of the extracted principle. The possibility of seduction only enters our consideration at all once we are brought to think of the second situation. Consequently, it is impossible for the Generalist to incorporate this situation by adding the phrase "Unless the person is trying to seduce the other" to the extracted principle. There are simply too many exceptions that would have to be be included, to the point where, if indeed these exceptions are all included, the principle loses its determinacy.

Further, while Dancy does not make this explicit in the text referred, there is a sense in which even the man's individual actions to make amends in the first situation becomes morally pernicious in the second situation. Each of the man's actions may change in moral valence as we start to consider the overall goodness of the situation as a whole by acquiring

---

[49] Ibid.

new information about the context. His actions that we would have considered as contributing to our moral approval in the first situation has the opposite effect on the second situation because we have new information about the man's intentions. E.g., payment for special care for the sake of the woman becomes a "necessary investment" for the man's seductive projects even when it provides support for the woman. These efforts add to our disapproval of the man's actions rather than the approval of the man's actions. The more elaborate and time-consuming the man's effort to provide care for the woman, the more approving or repulsive our reaction grows depending on the context. This flux in the moral valence of individual factors that constitute reasons for our judgment of the overall situation, is accounted by the holism of moral reasons. This is at the heart of the Particularist opposition against the sole reliance of moral principles, or SGP. The power of Dancy's argument against Hare's conception of "relevantly similar" situations lies not only in the internal weakness of Hare's notion of the "relevantly similar," but in the intuitive appeal of the holism of moral reasons. It seems simply implausible, once we consider these kinds of cases, that situations are generalizable to the point where a moral imperative drawn from one instance is reliably applicable to another, when the reason for us choosing one moral imperative is not guaranteed to have the same meaning for us in other situations. Picking out an individual factor that counted for our approving judgment, and saying that that individual factor should positively count for our approving judgment in a different situation sounds implausible once we become aware of this constant holistic flux.

* * * *

Moral principles cannot by themselves provide the discretion to determine when they are required. For, even if moral life is interpreted from the viewpoint that people hold certain moral principles (simply because people are instructed since our youth to not do certain things), and while people can generally follow their purported moral principles (meaning, one need not be a moral saint to be considered as faithful to moral principles), it must remain true that making exceptions for those principles in favor of another is as much an integral part of our moral life as remaining faithful to our principles. Applying moral principles to situations would require some ability to "read the room," as it were, and moral principles themselves cannot provide this ability. Even the choice of moral principles requires particular judgments

based on what is at issue.

But I should first establish that this ability is indeed required. We generally agree that murdering another is bad, and perhaps this is a good candidate for a moral principle that we could consider as being close to universal. But among this general agreement, quite a lot of people who agree to this principle would also be willing to accept that sentencing, say, a murderer the death penalty is acceptable (given that retaining the Death Penalty is still a contentious topic). And clearly the death penalty is willful killing of the other, regardless of the legitimacy of its motivation, and should look like murder as well. So, apparently, many people are willing to make exceptions for at least one moral principle. What this also reveals is the distinguishing of legitimacy behind actions that seem, in face value, morally unacceptable. The death penalty is seen as legitimate practice of the law while other killings are seen as illegitimate.

Note that there is no noticeable change, if any, in the moral valence of the act of murder in comparing the death penalty and other killings, in contrast to the case with the woman in the hospital suggested by Dancy. This distinction should be made because some may object that it is inconsistent to support the death penalty while saying that murder is bad. These people would argue that, for death penalty supported to be consistent, in stating a principle that "murdering another is bad," the death penalty supporters should implicitly attach "except in cases of the death penalty" at the end. But I think this need not be the case. In supporting the execution of someone, the act of murder itself is not understood as being contextually approvable—no matter who is being murdered, the act of murder itself is not being morally approved. It is simply a policy-wise necessity that the criminal be murdered, and in this sense, there is an exception-making to a moral principle rather than a change in the moral valence of the act of murder. In fact, if the death penalty is understood as the maximal punishment, it should indeed remain the case that the act of killing itself remain the most evil event that can happen for an individual. The attachment of "except in cases of the death penalty," which the objection claims to occur *concurrently* with the formulation of the principle "murdering another is bad," can only be reactive to the recognition that the principle itself should be suspended due to a given situation, and therefore cannot have been part of the moral principle in its inception. Of course, this unclear possibility of suspending the principle cannot itself be part of the principle. This is evident once we try an example: "Lying is bad unless you are playing Diplomacy and unless this principle does not apply." The inclusion of

an exception of playing Diplomacy is deliberate; adding concrete exceptions does not eliminate the possibility that this principle itself can be suspended in cases that cannot be legitimately expected at the time of the formation of this principle.

The Generalist may interject here and say that what occurs is in fact an *overriding* of one principle that is superior compared to another, and that this really only implies the existence of a higher-order principle to which all these replacements could be accounted for. For the death penalty example, the rule that replaces the "no murder" rule is perhaps that we should prevent harm done by criminals if we can. One could then argue that once this process of abstraction is done for a good number of times, then we could perhaps find a principle that could "rule them all." Or, we could have an infinite regress of principles and super-principles—the question is open, is it not?

We may stop here for now. To borrow Dancy's words here, we may wonder "why we should insist that some such defense is there to be found"[50] for SGP that is held by "Ready-Ethics." What motivates "Ready-Ethics" to defend SGP? We understand that SGP is the answer that "Ready-Ethics" gives against the normative question, but why SGP specifically, and how is the presentation of this answer made possible? Since we have now seen that SGP is at least a contentious, if not questionable, view, it would be useful to shed light on what motivates "Ready-Ethics" to hold this view as a complete and exhaustive construal of the fundamental structure of the moral life.

2.2.2 Instrumentalism of the Ethical in "Ready-Ethics"

"Ready-Ethics" expects Ethics to be postured in a way that enables the furthering of technological development—i.e., "Ready-Ethics" is Ethics postured in the form of a standing-reserve to respond to enframing, like a soldier waiting for morning roll. Surprisingly, Ethics taking the posture of "Ready-Ethics" is not usually found immediately objectionable. In fact, "Ready-Ethics" is specifically designed to be approved by our ethical screening. As we saw from Powers' definition of machine ethics, the goal of "Ready-Ethics" is to find moral acceptability. By treating the ethical as instrumental, contrary to what one might expect from the connotation of the word "instrumental," "Ready-Ethics" does not present an obviously untenable moral theory that *only* caters to the needs of the technological. In fact, it is the

[50] Dancy, p.81.

utmost priority of "Ready-Ethics" to avoid being seen as demoting the ethical. What "Ready-Ethics" does instead is to take the shape of a plausible moral theory that can most *enable* the development of artificial agents. We observed from 2.2.1-A that "Ready-Ethics" takes on the form of SGP, which is a crude version of Moral Generalism, one of the major contenders of moral theories in Ethics. But the technical advantage and the background upon which SGP was selected has not yet been explained adequately. In what sense is SGP the "most enabling"?

As my cited examples so far suggest, the main appeal of SGP is that exceptionless principles seem as if they are easily translatable into "formal specifications." Recall the "programming" and "training" approaches in Sharkey's account of Ethical AI development. Research by Winfield et al., which is an example of the "programming" approach, due to its clear portrayal of their architecture of the "ethical robot," sufficiently reveals the advantages for selecting SGP. The most important piece in Winfield's decision-making model of ethical robots is the Consequence Engine (abbv. CE), which provides the capability to "generate and test *what-if* hypotheses."[51] The testing of hypothetical situations and actions is done through pre-programmed principles. Having formalizable moral principles at hand would surely streamline this process. Further, it would also be profitable for the engineer if the factors involved in the evaluated situation do not change while different principles are being applied (*contra* holism of moral reasons, which we went through above as the main appeal of Moral Particularism). Hence, the pragmatic of SGP, *the validity and value of generalizable and exceptionless moral principles*, is clear for the programming approach.

Some may wonder whether only rudimentary models like Winfield's approach, which are meant to serve as "proof of principle,"[52] and are designed with no intention to be used as an immediately applicable model in the real world, would necessarily commit the researchers to embrace SGP. In other words, some could object that sophistication of the architecture could allow researchers to not adopt SGP. Indeed, this might be the case. And it is fortunate for us that the "training" approach is a good candidate for an improved architecture for us to consider. The "training" approach does not immediately give off the

---

[51] Winfield, A. F. , Blum, C. and Liu, W. (2014) Towards an ethical robot: Internal models, consequences and ethical action selection. In: Mistry, M. , Leonardis, Aleš, Witkowski, M. and Melhuish, C. , eds. Advances in Autonomous Robotics Systems: Proceedings of the 15th Annual Conference, TAROS 2014, Birmingham, UK, 1-3 September 2014. p.87

[52] Winfield, p.95

impression that it commits its researchers to embrace SGP. As Malle positions himself in his paper, this approach is one that grew in response to the perceived limits in the "programming" approach. Since the subtle rules and contexts involved in moral decision-making are too complex to be readily translated, Malle suggests that it is perhaps better to "train" artificial agents so that they acquire "moral competence."[53] However, the "training" approach as described by Malle reveals yet another aspect of "Ready-Ethics."

Before I go in to detail, however, I should clarify my intent in using the word "instrumental" to describe this additional aspect of "Ready-Ethics." There are two ways in which something can be instrumental. The first is to be useful for an end as a means. The end to which the instrumental is used for holds commanding authority over the form that the means takes. This is perhaps the more intuitive sense of the word. The second sense of "instrumental" is to think of an object as available as an *instrument.* Ethics poses ready as a standing-reserve, preparing to be used by the engineer. Ethics, like any physical object available in the world, is then treated as a resource that could be readily used for the development of the artificial agent. The usual impenetrability attributed to the normative question is ignored, and these questions are forced to be solvable. Ethics as the totality of decisions and considerations by humans become all *readily* applicable.

Through the explanation of these two senses, it should be clear how these are meant to be associated with "programming" and "training" approaches. The first sense of instrumentalism corresponds to the "programming" approach. "Ready-Ethics" provides a perceivably acceptable answer to the normative question (in the form of exceptionless principles) *only* so that technological progress is maximally enabled. The second sense of instrumentalism corresponds to the "training" approach. "Ready-Ethics" acts as if the normative question is already solved by us humans when to the normative question we can only give at best hypothetical answers, which, in turn, are only produced via the continuous critical consideration of particular situations which yields unpredictable results according to Particularism. This approach takes our hypotheses up to now as sufficient for shaping the normativity that will ultimately govern artificial agent decision-making. So, let us return to Malle again to see what the latter sense of instrumentalism exactly means, and how it distorts the normative question and errs in capturing the moral life.

---

[53] Malle, B.F. "Integrating robot ethics and machine morality: the study and design of moral competence in robots," *Ethics and Information Technology*. (4). p.246

In designing a scientific framework for use in understanding Ethics, whether or not this process involves Ethics' transmutation into "Ready-Ethics," one is liable to committing herself to a specific viewpoint towards the normative question with or without intention to do so. Malle does exactly this in his exposition of moral competence for artificial agents. In thinking of moral competence that should be required for artificial agents, he claims that "what we need to examine is not "true" moral competence, but the competences that people expect of each other."[54] This claim, while sensible and even prudent from the perspective of the technological (and above all, useful), is naïve from the perspective of the ethical and possibly even dangerous. After all, is it not the case that the moral competence that we gain through what is expected of us, and what *we ourselves* expect from us the result of always having the normative question—the idea of the "true moral competence"—in mind? To argue that the pursuit of these two can be separated is a stance that Malle takes regarding the normative question—he argues that moral life can be achieved even without striving for the ultimate good. Since Malle at times refers to our education of infants[55] to make his approach more intuitive, let me use the same example to clarify my point: is there no significant difference between educating a baby to make sure she does not cause any trouble and educating a baby, but instead with the intent to make sure she turns out to be as good a person she could possibly be? Surely, we do not educate our infants by teaching us what people usually do, but by showing them exemplary and exceptionally good cases of individuals to communicate the *true* sense of what *the good* and the *desirable* is, and not merely the *acceptable*. Why do we do this? It is because not all people are committed to being good and therefore are not fit to be models for our youth. Further, even those who are committed to being good make mistakes or make lapses of judgment. We even hold prejudiced, distorted judgments without us knowing of them being there until another person who is brave and kind enough to do so points them out for us. We only become morally acceptable, after our failures and mistakes are subtracted from who we strive to become as an ethical being. Ignoring this subtraction is exactly what Malle suggests by his claim that moral competence can be found through the observation of our own behavior—that in producing an ethical artificial agent, it suffices for us to refer to ourselves as *the* model. The example of Malle is an instance of the latter sense of instrumentalism in "Ready-Ethics:" "Ready-Ethics" claims

---

[54] Malle, p. 245.
[55] Malle, B.F., and Scheutz, M. "Moral competence in social robots," p.5

that how *we* live is an answer to the normative question—the good, the answer to the normative question, is treated as a readily accessible set of ready-made judgments that can be used for the development of technology.

Of course, Malle's idea is not the only instance of this latter sense of instrumentalism. Researches that use the voting-based approach to aggregate the opinions of people, and thereby find the appropriate action against a hypothetical situation, can also be understood as taking our actions as the final answer to the normative question.[56] In fact, these are the most egregious instances of this instrumentalism. While Malle's idea of moral competence may be freed from instrumentalism if its aim was reconfigured to "true" moral competence, these kinds of aggregative methods are not even equipped with that alternative possibility. That people's opinions, and their *aggregated* opinions, no less, provides the *right*, or even "better"[57] moral viewpoint is a requirement for this kind of research. Some may object here that I incorrectly insist that such aggregative methods are conducted with the assumption that opinions collected from the masses are the *right* opinions. Those who object in this way will say that they only wish to provide a guideline that lays out the "morally acceptable"—the goal of "Ready-Ethics." For these researchers to resist being portrayed as a "Ready-Ethics" practitioner, they must be committed to the view that these aggregated viewpoints will indeed end up with the *right* viewpoints. The danger of either of these views are clear to us by now: the former prioritizes the technological over the ethical, and the latter is too naïve to be taken seriously, and overlooks the fundamental corrigibility of our own moral judgments.

I chose to deliver these two different senses into the same label of "instrumentalism," to highlight the causality of these two by placing them in juxtaposition. Perhaps it would have been better to describe these two as committing the same crime of "demotion of the normative question." Among many ways to make an important question unimportant one is to make that question more accessible and easier to solve. What "Ready-Ethics" does with its instrumentalist treatment of the ethical, so as to circumvent the normative question, is precisely that: it suggests that the question was in fact an easy one to solve. By pointing to everyday human conduct as the readily accessible source for solution of the normative

---

[56] For instance, Awad, et al (2018).
[57] Conitzer, V., Walter, S., Borg, J., et al. "Moral Decision Making Frameworks for Artificial Intelligence." p.4834

question, “Ready-Ethics” destroys the normative question. There is no reason to wrestle over the “true” moral competence or “formalizable specifications” of ethical principles—the moral truths. It is sufficient to work with what is “available”[58], and the only remaining difficulty is the “awe-inspiring”[59] work of translating the moral opinion of the masses into code.

3. Moral Particularism in Machine Ethics

Ethical AI research so far was dominated by “Ready-Ethics,” which faced the inevitable task of answering the normative question. “Ready-Ethics” as enframing, the drive to organize things into a standing-reserve for the advancement of technology, presented SGP as the answer to the normative question. This answer-giving by “Ready-Ethics” was *enabling* due to the technical feasibility that would be allowed by exceptionless moral principles available in Ethics, if there were any available. The answer-giving, on the other hand, was forced *possible* through the demotion of the normative question into an empirical question, meaning that the normative question was made answerable by reducing it to a different one (to finding the generally acceptable) and thereby making it “discoverable,” either in exceptionless moral principles or in the everyday conduct of individuals. This what I argued so far.

The normative question in moral artificial agent design should be treated with the same kind of respect that we give when we deal with the normative question in other matters. This is not only because the normative question is our object of philosophical wonder due to its impossibility of being given definitive answers, but more importantly because our struggle with the normative question is essential to what it means to be human, and therefore, potentially moral. Being confident in front of the normative question is what we constantly strive for, even if there is no possibility of ever becoming so, and this effort to become confident therefore is what defines our conduct. Further, our conduct is in constant progress of development due to our disparities between the standards that determine what is defeasibly considered good, and the standards that determine what is good, all things considered. So, in designing an artificial moral agent, we should also continue to wrestle with the normative question, and not merely pretend that we do not have to deal with it as “Ready-Ethics”

---

[58] Noothigattu, R. et al., “A Voting-Based System for Ethical Decision Making,” p.1.
[59] Malle, B.F., and Scheutz, M. “Moral competence in social robots,” p.35

suggests. We should therefore also approach ethical artificial agent design with the following attitude towards the normative question: assume that we do not have immediately providable answers, and approach the matter with humility that necessarily results from trying to solve a problem that never fails to elude our grasp. However, with regards to our reaction to the normative question, what we do know, is the fact that we make decisions every day in moral situations, and that we do this despite lacking a *general* and *definitive* answer to the normative question at hand. We end up making choices in the end and require the means to do so. In light of this, we have seen that moral principles have fundamental flaws that makes it hard for us to accept SGP, which assumes that the moral life only consists of exceptionless moral principles. So, it should be that contextualism cannot be excluded from discussion of the moral decision-making model insofar as that model is devised by us humans, who do not have access to definitive answers to the normative question at hand. Therefore, if any convincing ethical artificial agent decision-making model is possible, then it should be the case that such unpredictable, context-sensitive considerations are incorporated as part of the agent's capabilities. Perhaps the first step in treating the normative question with respect, and therefore to "master" enframing, is to move away from the SGP approach in our design of moral artificial agents and consider the incorporation of some level of Moral Particularism.

Thus, we are brought to the following question, which I consider in this section: what does incorporating Moral Particularism into ethical artificial agent design really mean, and how would that look like in a technical application? What are the foreseeable problems with this approach? In this section, I explain how defeasible principles can play an important role in incorporating Particularism to moral artificial agent design. Then, I look into the possibilities of implementations explored by Guarini.

3.1. Defeasible Principles in Moral Decision-Making Models for Artificial Agents

Those who push most vigorously for the idea of defeasible principles in Ethics are perhaps Lance and Little. While they argue that their stance, which they label as "deep contextualism,"[60] stands apart from Moral Particularism in that they do not completely deny the moral significance of the use of generalizations in moral decision-making, their recognition of the limitations in practical application of exceptionless moral principles makes

[60] Lance and Little, p.54n5.

them thinkers that incorporate a certain level of Particularism. Lance and Little describe defeasible generalizations as found across many disciplines, ranging from "epistemology to biology, from ethics to semantics."[61] The notable characteristic of these generalizations is that, while they are "shot through with exceptions that cannot be usefully eliminated,"[62] "usefully" in the sense that these exceptions by themselves cannot be incorporated as part of the generalizations, they nevertheless serve valuable explanatory roles by being "rules of thumb or helpful heuristics."[63] What is fascinating about these generalizations is that, there is no statistically useful way to determine at approximately what "percentage," as it were, a regular occurrence becomes qualified as a defeasible generalization. Lance and Little provide these two generalizations as examples: "matches light when struck;" "fish eggs develop into fish."[64] Unless the match is wet, the former generalization proves successful. However, it is actually rarely the case that a single fish egg develops into fish, which is the reason why numerous eggs are produced numerously in "packs." Lance and Little's point is that defeasible generalizations are not used with the aims of consistent prediction about whether something will necessarily happen, but rather with the purpose of identifying what, in "conditions in which they *do* hold, are particularly revealing of that item's nature."[65] In other words, defeasible generalizations of an item reveal a condition of the item that is "*privileged"* in some sense when we think about the item.[66]

How does this idea of defeasibility apply to moral generalizations? For one thing, this idea is remarkably anti-SGP, because principles by themselves, under this idea, would not necessarily determine whether an action performed in a particular circumstance is good or bad. Lance and Little provides an example of lying, which is an action that certain Generalists, famously Kant, attempted to prove the unconditional immorality of. While these Generalist arguments are perhaps all well and good, and perhaps even perfectly sensible under the system in which those arguments are raised, we also have the clear intuition that is not just permitted, but even more desirable than telling the truth in board games such as "Diplomacy, in which lying is the point of the game," or when "confronted with the Nazi

---

[61] Lance and Little, p.61
[62] Ibid.
[63] Lance and Little, 53.
[64] Lance and Little, p.61
[65] Lance and Little, p.62
[66] Ibid.

concentration camp guards."[67] Lance and Little argue, and I think correctly so, that our intuition that in these instances we are allowed or even required to lie are correct, and that why this becomes so can be explained by understanding the moral generalization against lying as a defeasible one rather than an exceptionless one. Let us return to Lance and Little's example of Diplomacy, where the point of the game is to expand the territory of the country that you play as through, mainly, deception. It is clear that deception is permitted in this game only because there was a "prior agreement" that deception is permitted.[68] For such an agreement to take place, it also should have been the case that there was a shared understanding that lying is bad *generally*, or more precisely, defeasibly. The agreement that deception is allowed in diplomacy is justified by the pre-existing shared understanding of the status of lying, which is that it is "defeasibly-bad-making."[69] Lance and Little calls this relationship between the act of lying in Diplomacy, and how it is justified based on the defeasibility of lying as bad, as "justificatory dependence."[70] In the end, what we get from Lance and Little is a picture of a moral principle which only has *conditional* priority and validity depending on the situational judgment of the agent.

Marcello Guarini explored the potential of incorporating defeasible principles into ethical artificial agents, and as of March 2019 he is the only researcher who explicitly focused on the intersection between ethical artificial agents and Moral Particularism. While the focus of his research was to learn more about how our moral decision-making process works than to devise a Particularist framework of moral decision-making for artificial agents, I believe the implementation of neural networks that he used to simulate case classifications (which is what we indeed do when dealing with various situations somehow) reveals also the technical viability of a Particularist machine, and is therefore relevant for our current purpose, which is to explore the prospects of a Particularist approach to machine ethics that can drive us away from enframing-driven "Ready-Ethics." One thing that may be confusing as I discuss his work is that he argues that principles need not be considered as "exceptionless standards,"[71] and also thinks "contributory standards," which is conceptually equivalent to defeasible principles that Lance and Little endorse, are part of the Generalist camp and not

---

[67] Lance and Little, p.53
[68] Ibid.
[69] Lance and Little p.69
[70] Ibid.
[71] Guarini, M., "Particularism and Generalism: How AI can Help us Better Understand Moral Cognition," p.1

the Particularist camp. To clarify, Guarini considers three possibilities in which principles can play a role in moral decision-making.[72] The first is moral principles as exceptionless standards, which is what we have taken moral principles to mean so far. The second is moral principles as contributory (equivalent to defeasible), which is what Lance and Little suggest. The third is viewing moral principles as completely useless for moral decision-making, which Guarini calls the Eliminativist view, which is the view that none of the two kinds of principles exist nor are relevant in moral judgment. The third is the only view that Guarini attributes to Particularism, presumably because he wished to isolate the more radical view regarding principles to carry across his point.

To test which of these viewpoints are appropriate as a model of producing an artificial agent that could make correct moral judgments across multiple cases, Guarini set up one agent so that it is trained to output approval or disapproval based on four renditions of a moral predicament, and set up another agent so that it reclassifies the cases as either approved or disapproved based on the additional details that are added to the original predicament.[73] In his attempt to modify initial case classifications that were done in setting up the first agent via finding of what he calls *contrast cases*—similar cases with only one outstanding difference that is being tested as whether that factor counts for or against the approval the case[74]—he found that while contributory principle model could provide an intuitive explanation for how new cases could be reclassified based on the additional factors to cases, the Eliminativist model could not provide an intuitive explanation.[75] But the point about eliminativism is not what is of our interest now. What is important is that contributory principles, just like exceptionless principles, could provide reliable means for reclassification of cases based on new input. Reclassifying classified cases is perhaps one of the crucial aspects of human moral decision-making, e.g., we change our reaction to the car crash case given by Dancy. Additional research on contributory principle's potential to artificial agents may have great success in saving machine ethics from "Ready-Ethics" by drawing us away from SGP. But, will this Particularist direction lead to a moral artificial agent?

4. Moral Artifacts: Impossibility of Moral Artificial Agents

---

[72] Guarini, M. "Particularism and the Classification and Reclassification of Moral Cases," p.23
[73] Guarini, p.25.
[74] Ibid.
[75] Guarini, p.27.

Motivated by an attempt to move away from enframing-driven "Ready-Ethics," we have briefly considered the possibility of incorporation of Moral Particularism into ethical artificial agent design, so that some level of contextualism is incorporated in order to reflect the human condition in which the answer to the normative question is not readily available. Since Guarini is the only notable individual who has tried to specifically incorporate the Particularist points (by considering contextually shifting moral judgments) to the sphere of Ethical AI, there seems to be enormous potential to be gained in this direction.

But before I end the discussion, I consider what might await us in the end of the road. Not as technological fortune-telling, but as taking an idea and exploring its full implication. This may stand in contrast to the usual attitudes of AI researchers, for whom exploring the full possibility of technology is at times shunned and demoted to the area of science fiction, as the field was the target of over-excitement and great expectations many times since the 50s, and as a sort of a post-traumatic reaction, it has become prudent for experienced researchers in the field to take a reserved stance to the future to prevent such overreaction from the public.[76] But that is the stance that, perhaps, scientists should take. As for the philosophers, they are required to follow where the best arguments lead.

The full implication of the development of moral artificial agent is, in the end, a matter of not only artificial agents, but the inclusion and acceptance of new kinds of moral agents into our moral system. One may well ask, is this not yet another "Ready-Ethics" if we are talking about moral acceptability? The matter is, I think, different. If "Ready-Ethics" was focused on presenting a morally acceptable artificial agent to humans, in the sense that the responsibility on part of the *tool* to fit our moral standards, the question I wish to discuss here takes the hypothetical artificial agent as already having achieved "true" moral competence (preferably through the non-enframing approach), as it were. What we need to consider is what *our* reaction should be to that artificial agent. In other words, let us say we have an artificial agent that has achieved "true" moral competence, as far as we call tell from its moral judgments. A crucial part of being a moral agent is that the agent is a part of a community that the agent is willing to cooperate with, and that the agent does not only follow the norms of the community, but also is able to present his views so that they are, in the end, represented in the community's norms. For the emergence of an ethical artificial agent to be

[76] Conitzer, V. "Artificial Intelligence: Where's the philosophical scrutiny?"

truly possible, on top of the ability to produce valid moral judgments, people should be prepared to give some credit to the judgments of these artificial agents. For, what would be the use of an AI judge, for instance, if its verdicts that arouse the slightest of controversies are immediately discredited and deferred to a human judge? The point is not that artificial agents cannot be judges. Rather, moral judgments are already controversial with only humans, and it is likely not going to be the case that just because artificial agents are able to mimic moral decision-making, then all existing moral controversies will magically disappear. There will be disagreements to decisions presented by artificial agents. The point, therefore, is whether we can ever be prepared to give credit to such decisions made by artificial agents. For moral artificial agents to replace human moral agency where it is valuable to do so, it must be the case that the judgments of artificial agents are given credit, which would mean that their moral existence is accepted into the moral community. The question, in the end, is: Does the achievement of true moral competence on the part of the artificial agent necessitate that we accept it into our moral system?

Luciano Floridi, who is the initiator of what he calls Information Ethics (IE), presents a metaphysical system based on information[77] that is perhaps most open to this kind of acceptance. To the question above, he would answer with an emphatic "yes!" He is explicit in his rejection of the supposed "anthropocentrism" that dominates our current view of ethical agents:

> "… insisting on the necessarily *human-based nature* of the agent means undermining the possibility of understanding another major transformation in the ethical field, the appearance of artificial agents that are sufficiently informed, 'smart', autonomous and able to perform morally relevant actions independently of the humans who created them."[78]
>
> "IE is an ecological ethics that replaces *biocentrism* with *ontocentrism*. IE suggests that there is something even more elemental than life, namely *being* – that is, the existence and flourishing of all entities and their global environment – and something more fundamental than suffering, namely *entropy* … [which] here refers to any kind

[77] It is impossible to do justice to Floridi's grand idea of the infosphere in this space. A useful introduction of his metaphysical system for those who are interested can be found in *The Cambridge Handbook of Information and Computer Ethics*, which is included in the Bibliography.

[78] Floridi, L. "Information Ethics," *The Cambridge Handbook of Information and Computer Ethics*. p.87

of *destruction* or *corruption* of informational objects"[79]

That our understanding of Ethics should change if there are new possibilities of moral agents introduced does not necessarily sound like a claim driven by enframing, because without necessary regard to technological development and its production of artificial agents, there could be new possibilities of life (e.g., discovery of new tribes, extraterrestrial life forms with intelligence, human clones, etc.) forms that we find as acceptable into our moral community. However, if this argument was intended to focus on the inclusion of artifacts, or artificial agents that possess moral competence, I think one could give a very convincing argument that this idea is driven by enframing. In any case, I specifically consider the contention that "sufficiently informed" artificial agents must necessarily be included into our moral community.

What do I mean by inclusion into our moral community? I mean that the moral agent is treated as an equal to those who are already a part of the community, in every sense of the word. Inclusion means that the agent is not questioned in its capability to make judgments about the moral situation, and that its capability to give justifications for moral actions are not questioned as well. Some may wish to suggest that this kind of equality is not even achieved among humans, (e.g., epistemic injustice to oppressed groups), but I think there is still, at the moment of writing this paper, a fundamental difference in how people treat people and how people treat what we currently see as mere objects. Further, the people who treat other people as objects are considered as acting against the constitutive attitudes required of the moral community and therefore cannot be legitimately said to be a part of the moral community. And it is this gap between our treatment of humans and artifacts that I wish to discuss, because I believe it is the concern of Floridi's suggestion as well, when I discuss inclusion into the moral community.

Inclusion into the moral community poses a special kind of a concern to us, because to accept an artificial agent into the moral community is not like simply giving out a license of a sort to practice morality in our society, as it would be the case in the legal sector, for example, where the endowing of legal personhood is simply to "make arrangements that facilitate a particular set of social, economic, and legal relationships."[80] I think it is inevitable that we attach a certain sense of "personhood" to the artificial agent in accepting the future

---

[79] Floridi, p.84

[80] Chopra and White. *A Legal Theory for Autonomous Artificial Agents*. p.159.

moral judgments of artificial agents. For those who think the word "personhood" reeks with anthropocentrism, the sense I want to isolate is the strong fraternity and empathy we feel for the beings that we believe *deserve* better treatment compared to what they would have received under the objective conditions of nature without communal intervention. In attributing personhood, we give priorities in our treatment of certain beings that we accept to the moral community. For example, in the case of fire or emergencies or the like, we prioritize the rescue and evacuation of lives and not properties. What I want to conclude this paper with is that this motivation to prioritize those who have been accepted to the moral community is, in some ways, *necessarily anthropocentric*, which I will show that it is in fact sensible to maintain unlike what Floridi suggests, and that these grounds for anthropocentrism necessarily precludes the assimilation of artificial agents to necessary members in our moral community.

To clarify our task here, recall the "programming" and "training" approaches to designing artificial agents. We found that the programming approach, which requires exceptionless principles, is an enframing-driven approach because it leads to SGP. What we are left with, therefore, is the training approach, and we have established the viability of incorporating particularism into this approach by seeing the examples of Guarini where he experiments with defeasible principles in neural networks. The main goal of training approach is to achieve, as Malle claims, moral competence. So, we should consider whether the achievement of not only working, but "true" cognitive moral competence is enough for the artificial agent to be part of our moral community.

One question that we should ask here, is on *what justification* the moral competence of artificial agents would rely on. Habermas' distinction of three kinds of practical rationality provides insight to the justificatory process of moral decision-making that cannot be attributed to artificial agents unless at the same time *humanness* is attributed to these agents. Habermas argues there are "pragmatic, ethical, and moral employments of practical reason."[81] Of these, the latter two are in our interest. For the ethical, the question "Who am I, and what would I like to be?" is answered with the "unconditional imperatives such as the following: You must embark on a career that affords you the assurance that you are helping other people."[82] For the moral, to the same question asked in the ethical, the moral

[81] Habermas, p.9
[82] Habermas, p.5

community is called upon to answer "whether [the members of the moral community] all could will that anyone in [their] situation should act in accordance with the same maxim."[83] To clarify the distinction, the difference lies on the problem of how "one would want to live" and how this maxim "is suitable to regulate our communal existence."[84] "Ready-Ethics" makes a fundamental mistake by assuming moral competence as achievable through individual perfection of moral decision-making, when there is a crucial communality to morality as Habermas points out. What is good for the community, the Habermasian moral, can only be determined by placing himself and others in the situation where the law in question is universalized, and it is by this process of communal justification we get valid norms with "abstract universality" which the researchers of "Ready-Ethics" tried to apply directly to artificial agents.[85] The justifications of moral competence, therefore, cannot come from these norms themselves, but from the individuality and free will of the agent, who, with her own projects and interests, participates in the moral community. This presupposed self-interest in entering a moral community is what provides the legitimacy of moral justifications. Therefore, no mere mimicking of moral decision-making patterns on the part of artificial agents by itself could never be supplied with plausible justifications if the judgment is made by a being to which we do not attribute individuality. But this does not completely close out the issue at hand. After all, the view that machines will at some point reach and exceed human-level intelligence has considerable strength.[86] And perhaps to these human-level intelligent machines we could attribute also individuality. Insofar as I aim to justify the anthropocentric assumption in viewing artificial agents, I should also grant the possibility that artificial agents could achieve Artificial General Intelligence—that is, human intelligence—to test the viability of anthropocentrism.

Susan Wolf's consideration of moral saints proves very relevant in exploring this possibility. By moral saints, Wolf means "a person whose every action is as morally good as possible, a person, that is, who is as morally worthy as can be."[87] For our purposes, we can consider that an artificial agent achieved this status, if it is understood as having achieved "true" cognitive moral competence. Unless we accept the implementation of sometimes-

---

[83] Habermas, p.6
[84] Habermas, p.7
[85] Habermas, p.13
[86] *Superintelligence* by Nick Bostrom and *The Singularity is Near* by Ray Kurzweil both are well-read books about this possibility and its implications.
[87] Wolf, S. "Moral Saints," *The Journal of Philosophy*. p.419

moral artificial agents, which would be even more egregiously enframing-driven compared to the democratic aggregation of moral norms we observed above, we can safely assume that when we discuss morally competent agents in the context of accepting them as members into our moral community, we are talking about artificial agents at this level of moral competence of the moral saints. In any case, if we return to moral saints, we should recognize that "there is something odd about the idea of morality itself, or moral goodness, serving as the object of a dominant passion … [because] morality itself does not seem to be a suitable object of passion."[88] The point that Wolf wants to get across, is that there is a layer of selflessness that morality requires of us. However, there is also this layer of human nature where an individual has distinct, selfish interests that she wishes to achieve. In a way, to be a moral saint means to completely relinquish that aspect of self-interest and only focus on aiding the realizing the passions of other individuals. When a moral saint gives up his supposed favorite activity (Wolf gives "fishing trip" or "hot fudge sundae" as examples) for the sake of the duty that is imposed by morality, "one is apt to wonder not at how much he loves morality, but at how little he loves these other things."[89] There is an *inhuman* aspect to moral saints that make them uninteresting creatures. The point is similar to what Habermas makes. The saints do not have projects on their own and are only reflections of what the norms demand of them. This is quite implausible of a state for a human being to be, so one would be brought to question the happiness of the moral saint rather than be brought to admiration. One of the important, perhaps disturbing, point that Wolf makes is that achieving "true" moral competence goes against what we consider is essential to human nature, namely that of passion and self-interest, and that it would be undesirable and unadmirable to completely relinquish them for the sake of perfection in one's alignment with moral duties.

Wolf does not explicitly consider the case of artificial agents. So, as we bring her ideas into machine ethics and assess whether they can acceptably replace human moral agency, we are on our own. What is undeniable in Wolf's argument is that if we have a perfectly moral being, the being would perfectly reflect the requirements of societal norms, and so will be a being that does not have individual projects of its own. It follows, then, that artificial agents with "true" moral competence, even if they achieve human level intelligence, would be this kind of being. If artificial agents are only understood as tools while the

---

[88] Wolf, p.424
[89] Ibid.

achievement of moral perfection proceeds, this would not be problematic at all. In fact, that would be even desirable. But a problem takes form as soon as we consider accepting these kinds of morally saint-like artificial agents into the moral community, and as we consider whether to these artificial agents we should attribute the same kind of *personhood*, including the entitlement to correct us as fellow moral beings as we currently do to humans—in the end, that is what is required for moral artificial agents to legitimately replace human moral agents. We are brought to consider ultraintelligent, hypermoral machines as equals of the moral community, and that we should accept their moral judgments because they are "truly" morally competent than perhaps we humans will ever be. We have an intuitive repulsion to this notion—Floridi calls this unjustified anthropocentrism, and that there is no reason for us to be prejudiced against these additional life forms simply because they are non-biological, especially if they have proven their moral competence. Wolf's depiction of moral saints, however, shows that there are more reasonable grounds to this repulsion than mere prejudice in favor of the biological. The reason is that ethical artificial agents, if their implementation is possible at all, will be built with the expectation that they be morally perfect beings, while humans are not subject to the same kind of expectations when they only contingently exist. *By the application of this stricter ethical standard onto the creation of artificial agents, they are not brought into the world with the same level of freedom as humans, which is the crucial requirement for one to be part of the moral community as a project-having, free-willed individual.* They are not moral agents at all but merely artifacts that can produce moral expressions, because by being born as perfect moral agents they cannot said to individuals even if they have human intelligence, for there is no individual-community tension of interests to be found. For this reason, it is inconsistent to claim that one promotes a non-biocentric view towards moral artifacts as Floridi does, if, on the one hand he argues for the construction of ethical artifacts that have "true" moral competence (and he can only do so to move away from enframing-driven "Ready-Ethics"), while, on the other hand, it remains true that the ethical design of artifacts disqualify them as moral agents. Authentic non-biocentrism would have to argue that these restrictions on the freedom of artificial agents must be removed by giving up ethical artificial agent design—which is evidently nonsensical and dangerous. Therefore, the anthropocentric position, where only humans are reserved moral agency, is the only remaining sensible position in the matter of moral artifacts.

5. Conclusion

Achievement of true moral competence in moral artifacts, if possible, is only possible by drawing the direction of research away from enframing, and by incorporating Moral Particularism into the artifact's moral decision-making process. But an artifact's true moral competence does not translate into artificial moral agency that can replace human moral agency, because artifacts, due to the requirements of ethical design, do not share the conditions for having moral agency that allows them to make moral judgments whose justification will be accepted by fellow members of the moral community. That being said, this paper reaches no conclusions regarding the *agency* of AI itself. I only argue that they cannot be moral agents. To deny other kinds of influences that AI could have on different aspects of the world would be stretching my point too far. Indeed, in non-biocentric metaphysical frameworks (e.g., Floridi), it could be argued that artifacts have agency in certain respects. However, in any sensible metaphysical framework, attribution of moral agency to AI would be mere fetishism.

Bibliography

Anderson, M., Anderson S., Armen, C. MedEthEx: A prototype medical ethics advisor. In *Proceedings of the eighteenth conference on innovative applications of artificial intelligence*. Menlo Park, CA: AAAI Press, 2006.

Anderson, M et al. "A Value Driven Agent: Instantiation of a Case-Supported Principle-Based Behavior Paradigm." *AAAI Workshops*, 2017.

Awad, E., et al. "The Moral Machine Experiment." *Nature* vol. 563, 2018.

Chopra, S., and White, L.. *A Legal Theory for Autonomous Artificial Agents*. The University of Michigan Press: Ann Arbor, 2011.

Conitzer, V. "Artificial Intelligence: Where's the philosophical scrutiny?" *Prospect*, May 4, 2016.

Conitzer, V., Walter, S., Borg, J., et al. "Moral Decision-Making Frameworks for Artificial Intelligence," In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17) Senior Member / Blue Sky Track,* San Francisco, CA, USA, 2017, pp. 4831-4835.

Dworkin, G. "Theory, Practice, and Moral Reasoning." *The Oxford Handbook of Ethical Theory*, by David Copp, Oxford Univ. Pr., 2011, pp. 624–644.

Dancy, J. *Moral Reasons*. Blackwell, 1993.

Floridi, L., Cowls, J., Beltrametti, M. et al. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds & Machines* (2018) 28: 689. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L. *The Cambridge Handbook of Information and Computer Ethics*. Cambridge University Press, 2010.

Habermas, J. “On the Pragmatic, the Ethical, and the Moral Employments of Practical Reason.” *Justification and Application*. Cambridge, MA: MIT Press, 1993, pp.1-17.

Greene, Joshua et al, “Embedding Ethical Principles in Collective Decision Support Systems,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Goodall N.J. Machine Ethics and Automated Vehicles. In: Meyer G., Beiker S. (eds) Road Vehicle Automation. Lecture Notes in Mobility. Springer, Cham, 2014.

Guarini, M. “Particularism and Generalism: How AI can Help us Better Understand Moral Cognition,” in *Technical Report for Machine Ethics Symposium, American Association for Artificial Intelligence Fall Symposium*, Nov. 2005

Guarini, M. “Particularism and the Classification and Reclassification of Moral Cases,” *Intelligent Systems*, IEEE. 21, Vol. (4), 2006. pp. 22- 28.

Heidegger, M. “The Question Concerning Technology.” *The Question Concerning Technology and Other Essays*. Garland Publishing, 1977, pp. 3-36.

Noothigattu, R et al., “A Voting-Based System for Ethical Decision Making,” *Proceedings of Autonomous Agents and Artificial Intelligence (AAAI) Conference*, 2018.

Lance, M., Little, M. “From particularism to defeasibility in ethics,” *Challenging Moral Particularism*, edited by Mark Norris, Lance et al., Routledge, 2008, pp.54-74.

Malle, B.F. “Integrating robot ethics and machine morality: the study and design of moral competence in robots,” *Ethics and Information Technology*. 2015, Vol. 4. pp.243-256.

Malle, B.F., and Scheutz, M. “Moral competence in social robots,” IEEE international symposium on ethics in engineering, science, and technology. *Presented at the IEEE international symposium on ethics in engineering, science, and technology*. Chicago, IL:

IEEE, June 2014. pp. 30-35.

Schroeder, M. “A Matter of Principle.” *Noûs*, vol. 43, no. 3, 2009, pp.568-580.

Strahovnik, V. “Introduction: Challenging Moral Particularism.” *Challenging Moral Particularism*, edited by Mark Norris, Lance et al., Routledge, 2008, pp. 1-11.

Powers, T. M. “Models for Machine Ethics.” *Philosophy and Computers*, Vol. 14 No.1, p.4. URL= https://cdn.ymaws.com/www.apaonline.org/resource/collection/EADE8D52-8D02-4136-9A2A-729368501E43/ComputersV14n1.pdf

Powers, T.M. Prospects for a Kantian Machine, IEEE Intelligent Systems 21 (4), 2006.

Powers, T.M. “Prospects for a Smithian Machine,” International Association for Computing and Philosophy, College Park, Maryland, July 2013.

Wallach, W, and Colin A. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2010.

Wheeler, M., “Martin Heidegger,” *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), Edward N. Zalta (ed.), URL <https://plato.stanford.edu/archives/win2018/entries/heidegger/>.

Winfield, A. F., Blum, C. and Liu, W. (2014) Towards an ethical robot: Internal models, consequences and ethical action selection. In: Mistry, M., Leonardis, Aleš, Witkowski, M. and Melhuish, C., eds. Advances in Autonomous Robotics Systems: Proceedings of the 15th Annual Conference, TAROS 2014, Birmingham, UK, 1-3 September 2014, pp. 85-96.

Wolf, S. “Moral Saints,” *The Journal of Philosophy*, Vol. 79, No. 8, Aug. 1982, pp.419-439.