AI and Ethics (2021) 1:545–552
https://doi.org/10.1007/s43681-021-00070-3

ORIGINAL RESEARCH

A critique of the 'as–if' approach to machine ethics

Jun Kyung You[1]

Received: 21 April 2021 / Accepted: 5 June 2021 / Published online: 15 June 2021

Abstract
In this paper, I argue that the replication of the effect of ethical decision-making is insufficient for achieving functional morality in artificial moral agents (AMAs). This approach is named the "as–if" approach to machine ethics. I object to this approach on the grounds that the "as if" approach requires one to commit to substantive meta-ethical claims about morality that are at least unwarranted, and perhaps even wrong. To defend this claim, this paper does three things: 1. I explain Heidegger's Enframing [*Gestell*] and my notion of "Ready-Ethics," which, in combination, can hopefully provide a plausible account for the motivation behind the "as if" approach; 2. I go over specific examples of Ethical AI projects to show how the "as if" approach commits these projects to versions of moral generalism and moral naturalism. I then explain the flaws of the views that the "as if" approach necessitates, and suggest that they cannot account for the justificatory process crucial to human moral life. I explain how Habermas' account of the justificatory process could cast doubt on the picture of morality that the meta-ethical views of the "as if" approach proposes; 3. Finally, I defend the relevance of discussing these topics for the purpose of functional morality in AMAs.

Keywords Machine ethics · Ethics of artificial intelligence · Artificial moral agents · Philosophy of technology · Heidegger · Meta-ethics · Moral generalism · Moral naturalism · Functional morality · Habermas · Justification

1 Introduction

One of the objectives of the Moral Machine Experiment [2, p. 59] was to "contribute to developing global, socially acceptable principles for machine ethics." To do so, it gathered 40 million decisions from people by asking them to make choices on various versions of the trolley problem. The questions asked, for instance, to choose a person to save in a car accident if either a male or a female had to die. Then, the researchers identified patterns of moral preferences based on those decisions. Among their findings, the experiment found preferences for sparing the young over the old, preference for sparing the fit against the large, and preference for sparing the higher status over people with lower status. Needless to say, it would be terrible if such preferences were applied as principles that machines will "unerringly" follow [2, p. 61]. The experiment received significant attention but was also heavily criticized, most notably by Jaques [11]. This paper also aims to contribute to the prevention of ethically impoverished machine ethics experiments such as this. However, this paper goes further to suggest that the experiment's approach on morality itself, which I argue is shared by many other Ethical AI projects, was a part of the problem.

In Ethical AI projects like the Moral Machine Experiment, the aim commonly taken in replicating ethical decision-making is to create machines that look "as if" they perform ethical decision-making. Hence I name this the "as if" approach to replicating ethical decision-making. The most important supporters of this approach are Wallach and Allen [21], who, through their seminal work *Moral Machines*, proposed that machine morality should be built through dividable, intermediary steps that at first aim at the realization of "functional morality," as opposed to "true morality" that we take humans to have. This means that artificial moral agent (AMA) development should focus on achieving "functional equivalence of behavior" in ethical decision-making, which they claim "is all that can possibly matter for the practical issues of designing AMAs" [21, p. 68].

In this paper, I argue that the "as if" approach to ethical decision-making is insufficient for achieving functional morality in AMAs. I argue that we should not aim

✉ Jun Kyung You
junkyungyou@gmail.com

[1] Department of Philosophy, Tufts University, Somerville, MA, USA

to replicate ethical decision-making by only capturing the *effect* of ethical decision-making. I object to this approach on the grounds that the "as if" approach requires one to commit to substantive meta-ethical claims about morality that are at least unwarranted, and perhaps even wrong. Although the "as if" approach is advocated by Wallach and Allen and their followers on the grounds that we should dodge thorny issues in our understanding of Ethics by not assuming too much about morality, the "as if" approach, in fact, does not achieve this goal. Specifically, the precondition of adopting the "as if" approach is the acceptance of versions of Moral Generalism and Moral Naturalism. In other words, the "as if" approach assumes too much about morality for it to be a suitable approach to replicating ethical decision-making. I aim to suggest that the parochiality of the picture of morality adopted by these views can be a plausible reason for the ethical impoverishment of the Moral Machine Experiment. By doing so, I wish to launch a conversation of how morality should be viewed in replicating ethical decision-making.

To defend this claim, this paper does three things: 1. I explain Heidegger's Enframing [*Gestell*] and my notion of "Ready-Ethics," which, in combination, can hopefully provide a plausible account for the motivation behind the "as if" approach; 2. I go over specific examples of Ethical AI projects to show how the "as if" approach commits these projects to versions of moral generalism and moral naturalism. I then explain the flaws of the views that the "as if" approach necessitates, and suggest that they cannot account for the justificatory process crucial to human moral life. I explain how Habermas' account of the justificatory process could cast doubt on the picture of morality that the meta-ethical views of the "as if" approach proposes; 3. Finally, I defend the relevance of discussing these topics for the purpose of functional morality in AMAs.

At the end, the reader will see that this paper demotes a meta-ethical stance that has a privileged status in machine ethics. This paper is also a suggestion to start (or return to) talking about how morality is viewed in machine ethics. To start, I introduce Heidegger's concept of enframing which is proposed in his essay "The Question Concerning Technology." Then, it will be clear what I mean when I name the target stance as "Ready-Ethics."

2 Enframing

Enframing, the essence of modern technology according to Heidegger, "is a challenging that puts to nature the unreasonable demand that it supply energy that can be extracted and stored as such" [10, p. 14]. "Agriculture is now the mechanized food industry. Air is now set upon to yield nitrogen, the earth to yield ore, ore to yield uranium" [10, p. 15], etc. Technology's demand for energy commands and organizes nature into a "standing-reserve" [10, p. 17], or in a state where nature is prepared to have its energy extracted. This idea of the world becoming a standing-reserve is important for understanding enframing. Heidegger's discussion of the Rhine's transition into a "standing-reserve" is perhaps the clearest illustration of this:

> "The hydroelectric plant is set into the current of the Rhine. It sets the Rhine to 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 … 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" [10, p. 16][1]

His example is clearer once we think about the Rhine's purpose with regards to the bridge and the hydroelectric plant. Consider the Rhine with the wooden bridge over it. It seems funny to ask what the purpose of the Rhine is here. The Rhine is the landscape, and no more than an obstacle, if anything. Clearly, there is no purpose in the Rhine, at least in relation to the wooden bridge. Now consider the case where a hydroelectric plant is built into the Rhine. The Rhine is no longer a backdrop but a necessary part of the hydroelectric power complex. The Rhine supplies the flow of water necessary to turn the turbines in the complex, which produces power that will be stored somewhere, which will be distributed to homes and industries to, perhaps, power the lights in the streets or even a simple toaster. Each step of the process of harvesting, transferring, and using energy is purposeful, and carried out toward the "furthering of something else" [10, p. 15]. In sum, the Rhine is called upon, as it were, to have its energy extracted and is thereby organized by technology into a standing-reserve, i.e., into a state where it can have its energy purposefully extracted, stored, and distributed. By turning into a standing-reserve through enframing, the Rhine gains a new meaning (as a source of energy) with respect to the larger technological structure that extracts its energy.

[1] Italics mine.

Heidegger calls enframing unreasonable. Heidegger believes that the unreasonable aspect of enframing lies in the danger that enframing poses, as "it banishes man into that kind of revealing which is an ordering" [10, p. 27]. This means that enframing makes it impossible to access the truth of an object in the world (Heidegger uses the word "revealing," to emphasize the sense that the truth of the object comes to us, rather than the other way around). Accessing the truth of the object becomes impossible because we cannot avoid being exposed to the presentation of the object in the context of resource extraction (i.e., to the presentation where the object is "ordered" as a standing-reserve by enframing). Also, humans are "banished" in the sense that the object's additional meaning as a standing-reserve overshadows any other realities of this object. Humans are "banished" into a world where everything only exists as a standing-reserve. Heidegger claims that this is dangerous, because by making all objects lose their meaning other than that which they hold as standing-reserves, humans are also limited to seeing themselves in relation to a standing-reserve. As a result, they see themselves only as the "orderer of the standing-reserve" [10, p. 27]. In other words, the danger of enframing lies in the inability to access the "fundamental characteristic" [10, p. 27] of human essence without it being "marked" [10, p. 27] by enframing: humans can only exist as enforcers of enframing, or *human resources* whose function is to realize the transition of all things into standing-reserves.

In sum, enframing can be understood as presenting the world in a way where its objects are optimized to be useful for a further purpose. At the same time, enframing blocks other possible presentations of the world. Enframing is the process in which technology changes our relationship to the world by altering how the world is presented to us.

In bringing enframing into the discussion of machine ethics, I do not aim nor need to suggest that Heidegger was completely right about technology. Maybe that is true or false, but I do not venture to prove either. I think enframing is useful enough as an inspiration for considering the state of machine ethics, even if it ultimately fails to be a concept that is rigorously applicable to the state of machine ethics. As such, I only aim to suggest that the idea of enframing can provide helpful insights regarding the state of machine ethics.

That being said, how is enframing relevant to machine ethics? Heidegger suggests that we lose the ability to access a "fundamental characteristic" of our essence. I propose that Ethics, a human endeavor, loses a fundamental characteristic of its essence through the "as if" approach in machine ethics. Ethics turns into a standing-reserve through the "as if" approach. As Ethics is turned into a standing-reserve, Ethics is re-organized into a field that is more suitable for the propagation of AMAs by capitulating to certain metaethical views. I call Ethics that results from this alteration "Ready-Ethics."

3 Introducing ready-ethics

> "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 … "when ground-truth ethical principles are not available, we must use an 'approximation as agreed upon by society.'"" [15, p. 1].

Ready-Ethics is the reconstruction of morality through meta-ethical frameworks that most enable AMA development. It could be seen as a manifestation of enframing on Ethics, where Ethics becomes a standing-reserve and is made prepared for AMA development.

Ready-Ethics is motivated by the need for machine morality, and so is, at first glance, compelling. The typical reasoning for the need for AMAs goes as follows: AI is a tool that, in a sufficiently advanced state, should not only enhance, but completely replace humans where it is valuable to do so. For replacement to be as widespread as possible, the technology must be trustworthy. In AI, one of the requirements of trustworthiness is the artificial agent's fulfillment of appropriate norms without active supervision by a human controller. Evidently, then, ethical decision-making must be represented in the artificial agent in some way—a machine morality—to achieve trustworthiness. We therefore have the need for AMAs.

The trouble with replicating ethical decision-making on any level is that we have no decisive moral theory. In fact, controversies in meta-ethics that occur at the most fundamental level (e.g., on whether there are moral facts or not) suggest that we are a long way from having such a moral theory. It is no exaggeration to say that, when it comes to ethical decision-making, we are utterly confused about what it is that should be replicating.

Regardless, Ethics is expected to be answerable to the demands of technology to become a standing-reserve. Ethics is expected to be extractable, and have a "formal specification of ground-truth *ethical* principle" [15, p. 1] prepared for AMA research. This need is joined by the insight that, on the surface, the artificial agent does not need to actually perform ethical decision-making like humans do. In fact, an *appearance* of ethical decision-making seems sufficient for the purpose of engineering an AMA [21, p. 16]. We are therefore brought to choose the more feasible option. We are brought to use the best "approximation [of Ethics] agreed upon by society" [15, p. 1]. We must imitate the effects—the "competence" [16, p. 245], the "function," [22, p. 112] the

'as if'—of Ethics. It has become sensible to present the matter in such a way that there is no other viable choice, because machine morality is "inevitable," [21, p. 17; 18, p. 2][2] and we must find a way to mitigate the negative ethical impact that could be caused by them.

In this development, the appearance of Ethics takes the place of the substance of Ethics in the AMA development process, and hence in machine ethics. Here, the influence of enframing is gone unnoticed, where a "fundamental characteristic" of morality is lost in the extraction process. What the "as if" approach ends up with lacks this "fundamental characteristic" of morality: it is a completely different morality—morality that is reimagined by the pragmatic needs of AI development. It is a morality that has been made into a resource, where formally representable information has been made possible and readily accessible. It is Ethics that has been made *prepared* for Ethical AI. The "as if" approach ends up producing Ready-Ethics.

None of this origin story, if you will, shows what is problematic about this alleged Ready-Ethics. I also do not imagine that readers will find this story compelling on its own. The rest of the paper is devoted to making this account meaningful. To do so, I will first show that there is a sort of an identifiable meta-ethical trend that is caused by the "as if" approach to Ethics, which turns Ethics into Ready-Ethics. Then I explain how there are theoretical weaknesses to Ready-Ethics that may have functional consequences.

4 Ready-ethics: the metaethical obligations of the "as if" approach

> "First, we summarize global moral preferences. Second, we document individual variations in preferences, based on respondents' demographics. Third, we report cross-cultural ethical variation, and uncover three major clusters of countries … We discuss how these preferences can contribute to developing global, socially acceptable principles for machine ethics" [2, p. 1].
>
> "We consider approaches that query [people] about their judgments in individual examples, and then aggregate these judgments into a general policy" [3, p. 1].

Ready-Ethics, the result of the "as if" approach to replication of ethical decision-making, is a view *about* morality and therefore a meta-ethical view. It is a combination of versions of moral generalism and moral naturalism.

Moral generalism is understood as the belief that moral principles most accurately capture morality. For this paper, principles are understood as having absolute and contributory senses, where the former prescribes an action as universally wrong (e.g., deceiving is always bad, so I should not deceive the murderer at the door), while the latter contributes to the ethical valence of an action (e.g., I would be deceiving, but since I could save my friend's life by doing so, this is ultimately good) [4, Sect. 1].

Moral naturalism can be understood as a variation of Moral Realism, which states that there are "objective, mind-independent moral facts" [14, Introduction]. Moral naturalists, adding on this assertion, further claim that these moral facts are natural facts.

These descriptions are deliberately far from comprehensive. I offer the above definitions as conceptual "tags" or heuristics that I hope are useful to identifying the kind of meta-ethical commitments required by the "as if" approach. A conscious commitment of the authors to moral generalism or moral naturalism per se are not necessary. Rather, the point is that the way in which morality is conceptualized in Ready-Ethics assumes certain meta-ethical commitments that can be contested.[3]

First, regarding moral generalism, AI ethics' reliance on a principle-oriented approach to morality has already been pointed out and criticized [24]. Further, many representative works on machine morality either focus on the identification of moral principles that can guide artificial agents, or choose to implement such principles based on existing moral theories in a top–down fashion, which involves consideration of general principles to analyze concrete cases [9, p.7; 19]. Alan Winfield, Blum, and Liu's decision-making model of ethical robots is a particularly illustrative example of what a principle-oriented approach to morality in machine ethics looks like [23]. Winfield proposes the Consequence Engine, which provides the capability to "generate and test *what-if* hypotheses." The testing of hypothetical situations and actions is done through pre-programmed moral principles [23, p. 87].

That it has become almost a standard move in machine ethics papers to open with a brief mention of Asimov's laws of robotics[1],[4] only to have the laws' insufficiency justify the researcher's motivation to introduce sophisticated moral principles, also testifies to the prominence of the principle-oriented approach on morality. Often, the candidates are utilitarianism, Kantianism, and other deontological moral

[2] The word "inevitable" means different things for the authors of [18], and their different views are listed there. However, the common denominator of their views is their commitment that there is an identifiable necessity for AMAs, which is enough to motivate my view that this drives a particular conception of morality.

[3] I am grateful for the anonymous referee in pressing me to clarify this point.

[4] For instance, [20].

theories (e.g., [9, p. 7]), which depend on the meta-ethical commitment to the viability of generalizable moral principles. The discussion of possible implementations of AMAs are framed in terms of principles from the start.

Second, commitments to moral naturalism are most obviously found in projects that use preference aggregation to identify general policies that will govern artificial agents. Preference aggregation is known "as the aggregation of several individuals' preference rankings of two or more social alternatives into a single, collective preference ranking (or choice) over these alternatives" [13, Sect. 3]. In the context of Ethical AI development, this means that based on a consolidation of many instances of individual human preferences, a policy decision is made on how machines would operate. We have already seen an instance of preference aggregation via the Moral Machine Experiment. Preference aggregation must presuppose moral naturalism because it must treat the judgments of individuals as data that will be relevant to producing the effect of ethical decision-making. Individual judgments must be treated as data that are *available* through empirical observation, making them *accessible* natural facts. By accessible, I mean that the observable results of an action are construed to be useful in establishing moral facts. If we return to the Moral Machine Experiment for example, if the respondents chose to kill in a car accident the old over the young, then that collective response must be assumed as a moral fact, along with which comes the authority that binds us to act that way.

At this point, we can remove the conceptual "tags" we attached to Ready-Ethics. We can have more precise descriptions. Ready-Ethics holds that:

(1) Organizing the collectable data of our individual ethical judgments into patterns will yield principles that are sufficient for determining what to do;

(2) Such patterns and data, which are obtainable through observing human actions and judgments, are not only principles but accessible moral facts sufficient for determining what to do.

I think there are reasons to believe why these views about morality could be wrong. Because they can be wrong, it makes these views worth discussing about in the perspective of developing AMAs. The effects produced by AMAs will be determined on these meta-ethical views.

Consider the first commitment. Considering how moral principles can be insufficient for ethical decision-making is the way I argue against (1). An argument I find powerful against the efficacy of moral principles is the holism of moral reasons. The holism of moral reasons is that what counts as contributing to the assessment that an action is "good" might contribute to the assessment that an action is "bad" depending on the circumstance where the action is delivered. Dancy gives a useful account of this by suggesting an example involving a man who strikes a woman with his car and puts her in a hospital [5, p. 80]. It is apparent that we would 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 here 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. We should note that we cannot reasonably say that we approved of the first situation only because we *knew* that the man was not intending to seduce the woman, and that this knowledge must have been part of the original principle. This is because there could be a myriad of defeating reasons like these introduced to make the man's decision look repulsive. For example, the man may have provided such care only to prevent her from suing him and had no consideration for her well-being, and so on. If all the possible exceptions were indeed considered in the original principle, and if they were all spelled out, the principle would have little meaning, or it would be simply not feasible for us to refer to them in our decision-making.

The change of our overall moral judgment between the two situations is dependent on changes to the valence of the man's actions. The valence of each action is dependent on the information that is revealed to us. The man's actions to make amends in the first situation is not morally equivalent to what he does in the second situation, although the content of the action itself, if taken in isolation, does not change. For example, paying for the woman's care can be interpreted as a "necessary investment" if the man's seductive projects are revealed, as opposed to it being interpreted as a "thoughtful gesture." We should also note that in the seduction case, if the man were to put in more resource to pay for the woman's care, it would add to the disapproval of the man's actions rather than the approval of the man's actions. Depending on the surrounding context, the same action could have a completely different ethical meaning. So, the point that holism of reasons makes out of this is that reasons that support a judgment in a situation are not guaranteed to be applicable in other situations. There is no guarantee that the same reasons would work if the valence of actions constantly change depending on the circumstance. Therefore, generalized moral imperatives from situations cannot be reliably applied to different situations. This spells trouble for the sufficiency of moral principles, which puts the moral generalist in the defensive.

Now, consider the second commitment of Ready-Ethics, which is that an agent's ethical decision could be guided based on available data of individual moral judgments. Malle seems to make this point most directly, when he claims that "what we need to examine is not "true" moral competence, but the competences that people expect of each other" [16, p. 245]. According to Malle, moral competence that can be identified by examining how morality is currently practiced

by us should be the reference point in replicating ethical decision-making. The resulting functional moral competence stands opposed to, for instance, theoretical ideals of moral competence. What Malle suggests is that the knowledge of "true" moral competence is not necessary to producing morally acceptable actions.

But this claim could be questioned. After all, is it not the case that moral competence that we have gained is the result of always being concerned with the idea of the "true" moral competence in mind?

Malle and Scheutz refer to our education of infants to make his approach to moral competence more intuitive [17, p. 5], so I use the same example to further drive the question above: is there no significant difference between educating a baby to make sure she acts like others 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? In moral matters, it seems like a more sensible policy to educate infants not by teaching them what people usually do, and thus *acceptable*, but by showing them exceptionally good cases of individuals to communicate the sense of what is good and desirable. This is a more sensible policy not because those infants should be under the expectation to become like those exceptional paragons, but because those exceptional cases best show what is required by morality. To assume that moral facts relevant to ethical decisions can be identified through observable practices, I think, gives too favorable of an interpretation on how good people actually are. Although it is useful from an engineering standpoint to think that the currently observable conduct of people serves as a suitable reference point for other agents, it is naïve and even dangerous from an ethical standpoint to do so—as it would be if the results of the Moral Machine Experiment were immediately represented in autonomous vehicles.

The two views of Ready-Ethics constitute contestable assumptions on what is necessary for AMAs to perform ethical decision-making, and, effectively, assumptions on what morality is functionally about. These assumptions hint at where the "as if" approach misses the mark in replicating ethical decision-making.

Heidegger suggests that we lose a "fundamental characteristic" of our human nature as enframing limits the accessible meaning of objects to their function as standing-reserves. I propose that, similarly, what is lost in Ready-Ethics is the justificatory process in the moral life, which I think is essentially an individual's struggle with the *normative question* regarding their actions, where the individual finds that she should answer the following: "what should I do here? Is what I have done a good thing?" We struggle with these questions whenever they are raised because no answers appear to us as evident [7, p. 2], and so we do not even recognize the point where we can be confident of the answer we give to the normative question. But we also, regardless, find ourselves responsible for providing the best answer we can find to the normative question, which obliges us to the justificatory process. We are obliged to explain our reasoning behind our decisions.

What, then, is this justificatory process? Habermas provides a compelling account. Habermas argues that there are "ethical and moral employments of practical reason" [7, p. 9]. The ethical employment is the use of "unconditional imperatives such as the following: You must embark on a career that affords you the assurance that you are helping other people" [7, p. 5]. The moral employment is to ask "whether [the members of the moral community] all could will that anyone in [their] situation should act in accordance with the same maxim" [7, p. 6]. The distinction highlights the different motivations for justifying in the moral life: the former targets individual integrity while the latter targets communal harmony.

Based on this distinction, the Habermasian picture of moral discourse proposes that one's answer to the normative question is produced through the individual's decision to engage with the moral *and* ethical employments of practical reason. Morality, according to Habermas, is a negotiation between an individual and the community. The morally desirable can only be determined by placing herself 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" [7, p. 13] which the researchers of "Ready-Ethics" tried to apply directly to artificial agents. That moral competence is built on the individual-society negotiation implies that moral competence is not built based on one's adherence to norms, but is based on an individual's willingness to consider one's own values with the values of the community. In short, moral competence is not about following norms but engaging with it. This would lead to a very different picture of the plasticity of norms. Because the justificatory process is reliant on the interaction between agents with independent objectives that constantly negotiate, it also can explain how we revise our moral beliefs, or the way in which we apply our moral principles based on newly encountered situations. A moral fact, according to Habermas' model, does not exist in the static way that Ready-Ethics prefers.

Here, also, I do not necessarily claim, or must claim, that Habermas has everything right about the moral life. All I claim is that there seems to be additional layers to the moral life than what can be captured in terms of principles and moral facts. I took an example of the justificatory process detailed by Habermas to show what could also be important to how we make ethical decisions, but there could be more. I only intend to show that there is something lacking about the "as if" approach's treatment of morality. If the trustworthiness of AMAs is dependent on how we assess the actions of human agents, and if our assessments of human

actions are based on whether the agent amply engaged with the justificatory process, that Ready-Ethics cannot account for this aspect of morality is a functional weakness. Insofar as "machine morality is just as much about human decision making as about the philosophical and practical issues of implementing AMAs" [21, p. 8], replicating the justificatory process should be considered as a possibly necessary component for the practical implementation of AMAs.

5 Conclusion: defending the relevance

While I stand by the above objections, I do not expect that I can show that the views of Ready-Ethics that I argued against are wrong in one sweep. Further, I do not propose an alternative model for doing machine ethics in this paper. So, the primary intended effect of these objections is to show that discussing the meta-ethical commitments of the "as if" approach is appropriate and necessary by suggesting that the picture of morality it assumes can be contested. I also imply that the views of Ready-Ethics have not ascended to dominance in machine ethics because of how evidently true they are. I suggest that they were, rather, collaterals of an approach that have strong incentives from an engineering perspective. The "as if" approach is, in the end, an effort to re-create the observable effects of human agents' capacity of ethical decision-making, and not an effort to exactly replicate the *actual* way in which human agents perform ethical decision-making. It is simply pragmatic for the engineer to assume that there is an existence of useful patterns in our observable moral life [6, 12].

Therefore, I suggest that machine ethicists should start (or return to) talking about how to approach morality in the replication of ethical decision-making. The available literature on Ethical AI initiatives suggests that there are many who are skewed towards the views that I offered objections for, and the lack of examination of these views may imply that we could be missing some things about functional morality, insofar as functional morality consists in the appearance of the AMA to be able to produce actions that comply to our norms. Certainly, the approaches criticized in this paper as Ready-Ethics are not unchallenged [8].[5] But the existence of these challenges does not damage my point that there are problems with making an implementation–first approach to AMAs.

Accordingly, the objections to this paper I am most concerned with will not come from those who believe that the meta-ethical commitments of Ready-Ethics are true, or that I mischaracterize the views that these researchers imply through their projects. Rather, the objection I am most concerned with will come from those who believe that the discussions I proposed we should have do not possess significant importance from the practical perspective of replicating ethical decision-making. Indeed, I believe the validity of the "as if" approach is based on the thought that discussion of such meta-ethical commitments are irrelevant to realizing functional morality.

Specifically, one could object that the justificatory process that I point to does not have to matter from the perspective of implementing AMAs. They could say something like the following: if the data of ethical decision-making available is produced by human moral considerations, where is the need to figure out how human moral considerations are *actually* made when, from an implementational standpoint, we only need to replicate the result of those considerations that are displayed as judgments and actions? After all, the primary purpose of replicating ethical decision-making for AMAs is to gain trustworthiness from the users by showing the AMAs' compliance to norms, so that they can perform tasks for us with high autonomy, not so that AMAs exactly perform ethical decision-making like a human does, although that task itself could be interesting.

My reply is that such meta-ethical commitments implied by the "as if" approach matters insofar as such commitments would matter for human moral agents in determining the kind of actions that they would perform in the same situation. Some may squint at this response, wondering if I am equating AMAs and human moral agents. But I only mean that, in assessing the functional morality of an AMA based on the effect it produces, an effect of an AMA should not be seen as appropriate if they simply comply to the expectations we have about an action's appearance. They can only be ethically appropriate if they meet the expectations we have about the justification used to support that action. If AMAs are to be trustworthy, they should meet all the expectations we have surrounding the action of human agents. So, there must also be little doubt that an AMA produces an effect for the right reasons. The "as if" approach to morality, and its corresponding Ready-Ethics, seems insufficient to dispel this doubt because it cannot account for the consideration that occurs in the agent and among the community.

Further, we should also think that reasons must be provided for an AMA's actions because human involvement cannot be separated even from an "independent" action of an AMA. Even the level of autonomy that will be granted to such an AMA will be the result of a human being analyzing the task and giving them to an AMA. That an AMA's algorithms become too complex, for instance, cannot override the human responsibility involved. To think otherwise fogs too much of the involvement of humans in developing and placing the AMA, and leads to confusion. There is always the need to think through what reasoning an AMA

[5] I am grateful for the anonymous referee that pointed to this paper.

will arrive at a course of action because the actions of AMAs are extensions of our decisions.

In sum, my reply is that one of the required conditions of building a functionally moral machine is to make its decisions trustworthy to human agents. *The "as if" approach fails to fulfill this condition by not accounting for the justificatory process that is a combination of individual internal moral reasoning and communal negotiation of values, which is plausibly one of our standards for determining whether an action produced was a morally acceptable one.* We should therefore start (or return to) talk about how we should think about morality in machine ethics.

Of course, I describe an ideal course of action with regards to how we should consider morality in machine ethics. I do not make these arguments with the intent to undermine Ethical AI in general. I only claim that the problem of replicating ethical decision-making, even at a functional level, is more complicated than how some make it out to be. The "as if" approach aims to emulate the effects of moral decision-making. But only looking at the effect of morality makes the mistake of assuming too much about morality, and in turns fails to capture what seems essential to morality, even at the functional level. Examining the flaws of the meta-ethical commitments of Ready-Ethics was done with the hopes of showing this. Hopefully, these considerations could contribute to a better way of doing machine ethics.

Declaration

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Asimov, I.: Runaround. Astound Sci Fict **29**, 94–103 (1942)
2. Awad, E., et al.: The moral machine experiment. Nature (2018). https://doi.org/10.1038/s41586-018-0637-6
3. Conitzer, V., Zhang, H.: A PAC framework for aggregating agents judgments. AAAI (2019). https://doi.org/10.1609/aaai.v33i01.33012237
4. Dancy, J.: Moral particularism. In: Edward, N.Z. (ed.) The stanford encyclopedia of philosophy (Winter 2017 Edition). Accessed 11 June 2021
5. Dancy, J.: Moral reasons. Blackwell (1993)
6. Floridi, L., Cowls, J., Beltrametti, M., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**, 689 (2018). https://doi.org/10.1007/s11023-018-9482-5
7. Habermas, J.: On the pragmatic, the ethical, and the moral employments of practical reason. Justification and application, pp. 1–17. MIT Press, Cambridge (1993)
8. Gabriel, I.: Artificial intelligence, values, and alignment. Mind. Mach. **30**(3), 411–437 (2020)
9. Goodall, N.J.: Machine ethics and automated vehicles. In: Meyer, G., Beiker, S. (eds.) Road vehicle automation. Lecture notes in mobility. Springer, Cham (2014)
10. Heidegger, M.: The question concerning technology. The question concerning technology and other essays, pp. 3–36. Garland Publishing (1977)
11. Jaques, A.: Why the moral machine is a monster. University of Miami Law School: We Robot Conference. https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/MoralMachineMonster.pdf (2019). Accessed 11 June 2021
12. Jobin, A., Ienca, M., Vayena, E.: Artificial intelligence: the global landscape of ethics guidelines. Nat Mach Intell (2019). https://doi.org/10.1038/s42256-019-0088-2
13. List, C.: Social choice theory. In: Edward, N.Z. (ed.) The stanford encyclopedia of philosophy (Winter 2013 Edition). Accessed 11 June 2021
14. Lutz, M., Lenman, J.: Moral naturalism. In: Edward, N.Z . (ed.) The Stanford Encyclopedia of Philosophy (Fall 2018 Edition). Accessed 11 June 2021
15. Noothigattu, R., et al.: A voting-based system for ethical decision making. In: Proceedings of autonomous agents and artificial intelligence (AAAI) conference (2018). arXiv:1709.06692
16. Malle, B.F.: Integrating robot ethics and machine morality: the study and design of moral competence in robots. Ethics Inform Technol **4**, 243–256 (2015)
17. Malle, B.F., 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, pp. 30–35. IEEE, Chicago (2014)
18. Poulsen, A., Anderson, M., Anderson, S.L., Byford, B., Fossa, F., Neely, E.L., Rosas, A. and Winfield, A.: Responses to a critique of artificial moral agents (2019). arXiv:1903.07021
19. Powers, T.M.: Prospects for a Kantian machine. IEEE Intell Syst **21**(4), 46 (2006)
20. Sharkey, A.: Can we program or train robots to be good? Ethics Inform Technol (2017). https://doi.org/10.1007/s10676-017-9425-5
21. Wallach, W., Allen, C.: Moral machines: teaching robots right from wrong. Oxford University Press (2010)
22. Wallach, W., Allen, C.: Moral machines: contradiction in terms, or abdication of human responsibility?" https://www.researchgate.net/publication/257931212 p. 112 (2011). Accessed 11 June 2021
23. Winfield, A.F., Blum, C., Liu, W.: Towards an ethical robot: Internal models, consequences and ethical action selection. In: Melhuish, C., Mistry, M., Leonardis, A., Witkowski, A. (eds.) Advances in autonomous robotics systems: proceedings of the 15th annual conference, pp. 85–96. TAROS 2014, Birmingham (2014)
24. Whittlestone, J., et al.: The role and limits of principles in AI ethics: towards a focus on tensions. AIES (2019). https://doi.org/10.1145/3306618.3314289