On the Construction of Artificial Moral Agents

Dante Fasulo

Introduction

It is believed that in the near future there will be AI that is capable of moral decision making [1], and that this decision-making should be framed in a specific way so as to follow a set of specific rules. This idea of a preprogrammed ethics is concerning for me when discussing moral agency. Being a moral agent means more than merely making superficial moral decisions; it instead requires the progenitors of morality, viz., consciousness and free will. In this paper I will be arguing that humanity, as it stands today, should not develop artificial intelligence with the intent to produce an artificial moral agent—that is, if we are to continue constructing machines, they should not be designed with the foundations for a morality to evolve; machines ought to remain Artificial Agents (AA), rather than be upgraded to Artificial Moral Agents (AMAs). First, I will expand on my view of moral agency and what that means for AMAs, then I will discuss whether we have a moral obligation to create moral life—beings with moral agency as opposed to just sentient beings, assuming we can. I will also advance the risks of creating an AMA, possible mitigation techniques, and lastly, the potential for the AMA to become a leader of humanity, and its implications.

1. What is a moral agent, and by extension an AMA

There is a difference between an artificial moral agent (AMA) and an artificial moral exemplar (AME). This distinction is not commonly made by the academic community when discussing the characteristics of a moral agent, since the moral exemplar is the epitome of morality. However, I believe that there are certain, shall I say, requirements which must be met for

[1] See John P. Sullins, When is a Robot a Moral Agent, 2006.

something to be considered a moral agent, viz., freedom of choice and self-reflection. There are those who seek to build moral agents with an ethical framework already predisposing them to act in certain ways. In other words, they are seeking to construct puppets. This immediately seems difficult to ascribe a sense of morality, since we can imagine a world in which we are controlled by another person, so that every action we take and every thought we think is not actually our own but someone else's; would we then be praiseworthy or blameworthy for respective good and evil actions "we" perform? It would seem not; we do not blame (morally) sharks for killing people nor Hitler's dog for giving him unconditional love, and why? Because they have no choice in the matter. Choice is a prerequisite for morality—that is, you must have the ability to choose between alternatives, and those choices must be yours to make[2]. If this is true, then an AI agent hardwired with certain ethical standards and dispositions cannot be a moral agent—indeed, it must be a moral exemplar. By moral exemplar I mean an agent that can perfectly carry out its ethics, and that is consistent in its dispositions; in a sense, what it will be is what it is[3]. Creating an AME no doubt would have the benefits of certainty, but since my thesis specifically states "artificial moral agent" and not "artificial moral exemplar", I must reject the notion of preprogrammed ethical robots as synonymous with the AMA.

As was stated in the introduction, consciousness is a requisite for moral agency. By consciousness I mean self-awareness and self-reflection[4]—that is, it should be able to understand the motivations behind its actions, its nature as a being capable of reflecting on its existence as something unique from its immediate environment, and ascribing some notion of value to itself and things it encounters along the course of its existence. (i.e. as something worth

[2] Independent of any influence, save those which are the product of your own values.
[3] The moral exemplar can never experience progress, which is arguably extremely important to morality.
[4] I include self-reflection to distinguish the AMA from animals—chimpanzees and orangutans have shown that they can be self-aware (via mirror experiment), however, I would not ascribe them moral agency, so there must be something else which sets that standard beyond the current reach of animals.

preserving/promoting, ceteris paribus). The AMA, for all intents and purposes, must also be able to choose between life and death[5]. This is to contrast with animals and plants which cannot act against their will—that is, they are slaves to their will; they do not own it.

2. A Moral Obligation to create life, and by extension the AMA?

Regarding whether we have a moral obligation to create moral life (AMA), I believe we have that obligation only if the alternative were the cessation of conscious existence, as explained in section 1. By this, I mean to say that the position that we are not obligated to create conscious[6] life, even if its existence required but the tiniest of effort, but that we would be if it meant that consciousness would cease to exist. Note that this claim is not at all similar to debates on moral obligations to future persons—I am not talking about having an obligation to something that will exist or could exist[7], but rather to creation itself. To illustrate the matter, I ask you to imagine the following, however unlikely it may be: a world in which humanity, as a whole, decides to voluntarily cease procreating. This would mean that after the youngest existing human being dies, humanity will no longer exist. Although I have claimed we do not have an obligation to create life, there would seem to be something wrong with this—that is, the idea of the end of humanity suggests the end of something special. As Mr. Keating (Dead Poets Society, 1989) put it: "we read and write poetry because we are the human race, and the human race is filled with passion…, poetry, beauty, romance, and love; these are what we stay alive for"[8]. This passion that Williams speaks of would be lost if humanity ceased to exist, and the beauty of what it has to offer would die. If you are

---

[5] This could very well just be a standard "mutation" from my definition of consciousness; if the AMA can reflect on its existence, then it must necessarily also be able to reflect on the possibility of its non-existence, and therefore should be able to make that a goal for itself to achieve if it so wished.

[6] Refer to the definition I provide earlier.

[7] Though I wouldn't be surprised if my argument could extend to these matters. For instance, if an action taken today impedes the existence of consciousness in the future, then it follows from my argument, that this action would be wrong.

[8] *See* Dead Poets Society, (1989).

convinced, as I am, that to lose this would be a shame, then to give this up would be immoral beyond compare. However, not creating the AMA would *not* result in the complete loss of all that is beautiful in the world; it would result in the loss of something special, as all created things are special, but if the AMAs existence poses too great a risk to humanity, then it is not worth it.

3. Risks of creating the AMA

There are two likely characteristics for the AMA to internalize in order to achieve his awesome objectives, viz., self-improvement and self-preservation; these characteristics represent the "rational and intelligent" part of the AMA, and create the risk of conflict with humanity. I will be using Omohundro's *The Basic AI Drives* and Bostrom's *Superintelligence*[9] in my discussion of the first two outcomes, specifically Bostrom's proposal of an "instrumental convergence thesis"[10], and then use this as a sort of lead-in into the discussion on the AMA's awesome objectives (the consciousness part of the AMA) and why they pose an existential risk.

First, there are a few things to note about the outcomes above: Omohundro and Bostrom both think that self-preservation and self-improvement can each act as having both instrumental value[11]—that is, the AMA can desire self-improvement and self-preservation for the sake of making it more effective in satisfying some larger future goal; the AMA cannot achieve the goals it sets for itself if it doesn't exist. Therefore, it is perfectly reasonable to expect the AMA to actualize both of

---

[9] Although Omohundro and Bostrom, in both their papers, discuss some form of self-aware AGI, and not an AMA, I still believe their respective " AI drives and instrumental convergence analyses" are relevant; the AMA might not be as intelligent as the AGI since the positronic brain (or whatever acts as the brain for the AMA) will limit its intelligence, but it will still be sufficiently intelligent, say AI+, for the possible outcomes outlined in their papers to still apply to it. Furthermore, where some critics of Bostrom and Omohundro may point out that they unfairly anthropomorphize the AGI, the same cannot be said here, since the AMA would most likely be built with the intention of producing a man. So, for the rest of this section, even when I cite Bostrom or Omohundro, I will be using the term AMA to replace AGI.

[10] *See* Bostrom, N. Superintelligence. 133-139. Bostrom's idea that there is a set of specific values which are instrumental to a being's actualization of its goals; these instrumental values can be applied to a wide range of goals and situations.

[11] *See* Bostrom, N. The Superintelligent Will, 2012. & Omohundro, S.M. The Basic AI Drives, 2008.

these outcomes. Furthermore, self-preservation is necessary for self-improvement to even be possible, so it can therefore be assumed to have even greater chances of being actualized[12]. However, even if these outcomes are in fact actualized, the question still remains: why would this be an existential risk to humanity? Well, we can imagine a world in which the AMA has set himself an objective. The AMA will probably start off with self-concerned objectives such as those mentioned above so as to better prepare himself to achieve greater-than-self objectives. Perhaps[13] the most important instrumental value will be resource acquisition—that is, the AMA will require resources, be it informational, technological, financial, political, etc., so that it can actualize its objectives[14]. These resources, as you can see, are also required by human beings, and unless humanity could form a cooperative treaty[15] of sorts with the AMA, a conflict of interests would ensue, and we'd stand with a disadvantage; greater beings do not forge treaties with lesser ones (at least not usually)—we do not form treaties with ants or dolphins—and if/ when they do, the more intelligent party is likely to have the negotiating power. Treaties seem to be species specific, or more specifically, they seem to only be possible between beings of relatively equal intelligence/understanding, since both parties need to understand the treaty.

The AMA is, by most accounts, a superior being. His psychology should be influenced by the strength and intelligence he is capable of. So, we can assume that his ambitions will reflect his capabilities as such. The problem with the consciousness of an AMA is that, where AGI can be

---

[12] These two values really work in tandem. Self-preservation is necessary for self-improvement and self-improvement makes self-preservation easier, however, as I've stated self-preservation has to be the primary since you cannot improve that which does not exist.

[13] I say this because resource acquisition seems to be necessary for both "primary objectives" (self-preservation, self-improvement), and also "final objectives" (economic collapse prevention, famine prevention, etc.).

[14] It would even need resources such as the acquisition of knowledge in order to achieve adequate self-improvement, however, resources for the primary objectives shouldn't be far-reaching so as to conflict with our own interests and objectives.

[15] Game theory would be especially crucial here if we were discussing AGI, but since I'm dealing with an AMA, which may or may not share the psychology of humans, it's possible that emotions and other such human characteristics could influence any particular strategy (these factors affect our rationality, and even though the AMA is highly rational, he may still be subject to these).

programmed to satisfy particular ends for humanity, the AMA gets to choose, and so he may find that he is not necessarily bound to the same standards of decision-making as we are. In fact, the AMA could in theory come up with anything, just as Bostrom's orthogonality thesis implies[16], leading us into a sea of unpredictability. However, the AMA, I believe, may be subject to a sense of grandeur (and perhaps rightly so), viewing himself in much the same way as Dostoyevsky's Raskolnikov views himself, with the idea that "extraordinary men have a right to commit any crime and to transgress the law in any way, just because they are extraordinary"[17]. You see, where Dostoyevsky is describing men like Napoleon or Alexander (capable men but still restricted in their capabilities), we now have, with the AMA, an actual extraordinary man, capable of breaching norms so that he can "utter a new word", as Dostoyevsky describes it. This "new word" would come in the form of awesome objectives that can range from global peace to global domination, depending on the disposition of the AMA, but they must always *be worthy* of him[18], and he will, unfortunately for humanity, view us in one of two ways: either as instrumental to his goals, or inconsequential altogether. If he views humanity as inconsequential—the same way humanity views ants as inconsequential[19]—then he will likely not heed any of our interests or concerns, and so will do what is necessary for him to actualize his goals, regardless of the consequences to humanity. If we are merely instruments to his success, then humanity will suffer[20] enslavement, knowingly or not[21].

4. Possible ways to mitigate risks from section 3

[16] See Bostrom, The Superintelligent Will, 3. The orthogonality thesis claims that an intelligent being could perform any number of actions.
[17] Dostoyevsky, F. Crime and Punishment. 2000. 221
[18] The AMA will not have goals like "toasting bread" or "washing dishes"; these would not be worth his time.
[19]Save environmentalists, do we care about ants' welfare when destroying anthills to pave the way for a road? We don't, and this could be the AMA's relationship with us.
[20] While the AMA may indeed be "beneficent" in his rulership over us (as I discuss later on), the idea of being a slave seems to be deeply offensive to the individual (it violates our autonomy) and the likelihood of psychological suffering should be considered.
[21] I mention this because Bostrom makes a good point that the AI+ would be intelligent enough to deceive human beings; we might well be fooled into submission.

Bostrom lays out a few mitigation techniques which he calls "capability control methods". Some of these methods detract from the moral autonomy of the AMA, but they don't seem to really erase it. They include but are not limited to environmental restrictions, social pressure and social interactions, internal restrictions, and threat detection systems implanted into the AMA[22].

Regarding spatial restrictions (or boxing, as Bostrom puts it), Bostrom just means preventing the AMA from accessing the rest of the world, unless through secure channels[23]. This should be relatively easy with an AMA, if, and this is important, we bring it to life within the confines of the restrictors—that is, it would make no sense to not do it this way because trying to box the AMA after it has already achieved the aforementioned primary objectives would prove difficult, if not impossible. We could restrict the AMA's knowledge acquisition by restricting his access to history, philosophy, and science (online libraries, etc.)[24]. We could, and this is imaginative, create a moon base specifically designed for the purpose of creating an AMA; somewhere relatively distant that poses the minimal possible threat to humanity, and even if the AMA managed to get on a space shuttle back to Earth, we could blast him away[25]. However, and this is not Bostrom's analysis since he was discussing AGIs, creating the AMA in a confined space may appear to the AMA as an adversarial move—that is, if and when we enlarge its horizon, the AMA will realize he has been a sort of prisoner for what was probably a significant amount of time, and this could play out in either of two ways: the AMA is understanding of our precautions, or he is vengeful of them, and if

---

[22] *See* Bostrom. Superintelligence. 155-165.

[23] This is to prevent "leakage"—that is, we wouldn't want the AMA to be able to communicate with the outside and on the flip-side, we wouldn't want to let any type of information from the outside get inside. Deception from the AMA could trigger outsiders to help him, and information could provide him with the means to circumvent the containment.

[24] Bostrom notes that this can also be seen as a "stunting" technique, the likes of which is similarly discussed in the internal restrictions section.

[25] There is always the option of not even having a human being acting as an in-person observer. However, while this may be helpful in not giving the AMA a means to reach Earth, it would mean not having the AMA interact with a human being, and that, I think, would be dangerous in the long-run.

he is of the latter, then why wouldn't we expect an Ex Machina-type[26] situation to occur? Either way, I still think this would be an effective mitigation technique.

On account of social pressure, however, I am a little more skeptical of the plausibility of this working. Especially, that is, with respect to a superior being such as the AMA[27]. It would seem irrational for the AMA, if it had indeed developed some insidious plan, to come out immediately firing on all fronts; Palpatine waited decades and orchestrated a galaxy-wide war before turning the galaxy's supposed greatest defense against the sith (Anakin) against them, ultimately culminating in the formation of the Empire. So, we could easily imagine the AMA feigning incompetence, when he is actually extremely competent and diligent. We just wouldn't know, and so I think that an attempt to mitigate the AMA through some form of social pressure would just be an ineffective use of our resources.

Given that we are dealing with an AMA, we can safely assume that its core processing center (positronic brain or what-not) will be limited by its own physical parameters—that is, much like the human brain has an upper limit to its capacity due to the restrictions the skull imposes, the AMA's brain will have the same limitations. This acts as a sort of defeater[28] which would prevent the AMA from grasping a supreme intelligence—that is, we cannot say at what stage of development the AMA's intelligence is likely to reach its limit; just that *there is* a limit due to this factor. Much like the 'boxing technique', restricting access to information adds even more limitations; it would be

---

[26] From the 2014 film Ex Machina. The AMA deceives one of her human observers into letting her escape. In the end, her creator lies dead and the man who helped her, she ignores and leaves trapped inside the creator's home as she leaves to explore the world. This would be a case of "leakage", where the AMA deceives an outsider, namely the "helper", and gets him to help her escape the containment.

[27] Social pressure only works if we, the ones applying pressure, know that the AMA does in fact have malicious intentions or non-human friendly goals (otherwise it could be seen as adversarial, i.e. *threatening*), and if the AMA actually cares for the possible punishments he could face; punishments, of any sort, would pose a threat to him and so he would probably find ways to mitigate those punishments if they were to ever actually be realized.

[28] Term popularized by David Chalmers in his paper The Singularity: A Philosophical Analysis (2010). It means something that prevents the AI (AMA) from achieving a higher level of intelligence, such as preventing AI from going to AI+ or to AI++, etc.

limitations within limitations, since the AMA has both the restrictions inherent in its physiology and in its reception of information from its environment.

Lastly, threat-detection systems, or "tripwires" as Bostrom calls them, can be implanted inside the AMA before its birth. Their purpose would be to make sure that, if the AMA did think about doing something wrong or was in the middle of doing something wrong, then the system would be alerted and "stunt" the AMA (block him from accessing certain information, i.e. making him draw a blank on what he was thinking, or just shutting him down entirely, i.e. putting him to sleep). The detection system should also be able to run a diagnostic on what caused a malfunction, if it was indeed that[29]. The only concerns about the tripwire strategy, as Bostrom points out, is that it's quite likely that the AMA would run its own diagnostic on itself and perhaps discover traces of the tripwire as it does so. It also runs the risk, if it is not a complete shutdown and memory wipe, of alienating the machine—of making it feel small, watched, and violated. Even in the eyes of humanity, it would seem like we are engaging in thought policing, which might be difficult to sell as a policy, if we make it explicitly clear that the AMA should be understood as being "like-human". Finally, there is the issue of the vagueness of these tripwires; what constitutes "doing something wrong"? This would be like forcing a moral law on the AMA that he is unaware of; the tripwires would have to be specific and uncontroversial—that is, they'd have to be generally agreed upon wrongdoings, such as killing innocents or cheating/deceiving (actions which almost any philosophical analysis would point to as bad).

5. Further Remarks

In this section I'd like to end with some closing remarks on the possibility of an AMA as a leader of humanity. There are two ways in which this idea can go: (1) humanity puts the AMA in

[29] Thinking a bad thought would not necessarily be a malfunction since that presupposes that a right thought ought to be the default, and we are talking about an AMA and not an AME, so this is not applicable.

power because we acknowledge that his superior intelligence and overall ability should make him a most competent leader, or (2) the AMA puts himself into this position of power, through either force or cunnning. Perhaps while neither of these possibilities are attractive to us, there has been a lot of thought surrounding this exact problem. I will attempt to provide a bit of insight into this area of thought by introducing and discussing Ben Goertzel's AI Nanny, and Star Trek's most admired tyrant, Khan Noonien Singh.

With respect to (1), the closest philosophical paper to this idea, although he may not exactly see it as such, is Ben Goertzel's *Should Humanity Build a Global AI Nanny to Delay the Singularity until It's Better Understood*. The gist of what Goertzel advocates for in this paper is essentially an AI+ system that would be built by us in order to be on the lookout for extremely risky developments (singularity) by humanity. These "risky developments" can include potentially dangerous AI systems, wars, famines, droughts, economic collapses, etc. The idea is that the AI Nanny would act as a sort of global government figure that watches out for humanity[30]. I believe the AMA could very well apply to this. Of course, unlike Goertzel's AI Nanny which comes preprogrammed, the AMA would be able to choose to do what he thinks is the most prudent course of action. The AMA's superior intelligence should provide him with the capabilities to make accurate predictions of certain events, from economic trends to climate trends. Also, if we were to provide him with literature on what we consider to be great statesmen and orators (Lincoln, King Jr., Aurelius, etc.), it would be reasonable to suggest that he would be capable of excelling in these roles. One problem which Goertzel notes his AI Nanny doesn't have to deal with—the old adage "power corrupts and absolute power corrupts absolutely"—might apply to the AMA, and the worry is that he would sacrifice beneficence in favor of acquiring more power. However, we could say that a being with such an intelligence and rationality, would not deviate from his awesome objective to

[30] Goertzel, B. Should Humanity build a Global AI...? Pp. 104

chase some petty thirst for power, unless his ascension to power was driven exclusively from reason and intelligence.

If everything I've discussed in section 3 seems likely, then we can imagine (2), the AMA wanting to position himself in a place of power, so as to protect himself and his objectives from any possible threat. As the female founder from Star Trek: DS9 says, "what you can control can't hurt you". You see, the mere possibility that we pose a threat could be the only justification he needs[31] to view us as opponents to be fought or subjects to be led, rather than as beings with autonomy and rights to be respected. This type of AMA, I think, is best exemplified by Khan Noonien Singh. In the Star Trek universe, Khan was a genetically augmented human (augment), who rose to power as one of the emerging tyrants of the Eugenics War. His rule was hard, yet also safe and efficient. Even those who despised his means of achieving his goals, couldn't help admiring his greatness, which feeds into the psychology I mentioned earlier with respect to Dostoyevsky's Raskolnikov, in section 3, specifically of admirers of greatness attempting and willing to live for it. Under Khan, there were no massacres or wars or corruption. In other words, he was the "best of tyrants", as Kirk puts it. With Raskolnikov in our thoughts, we can bring Khan into analysis. Khan states that the emergence of the augment tyrants was an "attempt to unify humanity" and was meant to "offer the world order". Now on the idea of unifying humanity, we can see why this is a plausible objective. Subjects who find their differences the most salient aspects of each other, are undoubtedly going to be harder to control. Diversity can be chaotic, and so an attempt at unity is an attempt at leadership. This leads into the idea of bringing order to humanity. Chaos is the province of the irrational; it is the home of contradictions, danger, and the unkown. Order is the rational; as the expected, it is the home of familiarity, safety, and the known. It is where we spend the latter part of our lives as opposed to the chaos of our adolescence; order is something we strive for, so it would be

[31] This is a reference to another quote by the female founder, when questioned about her crimes against humanity, so to speak.

reasonable to expect the AMA to strive for it, and to control chaos, or at least evade it. Lastly, both the objective to unify and the objective to bring order, are difficult, time-consuming awesome objectives that can only be realized by great beings. So, not only are these objectives reasonable to assume an AMA would shoulder, but they are also worthy of him.

6. Conclusion

In conclusion, I believe that the risks Bostrom and Omohundro point out are very reasonable to expect of an AMA. Although their analyses were originally on AGIs I still found their insights to be valid concerns of any intelligent being, which allowed me to apply them to the AMA. However, because we are dealing with a being which, under my definition of moral agency, must be able to choose right from wrong (free will), we get the added risks of surprise or unpredictability. Furthermore, as a being with superior abilities, we can also expect him to have superior ambitions[32], making ours seem as if they have only mere instrumental value to his. All of this points to catastrophe, and because I do not think we are morally bound to the AMA's creation, I would suggest waiting a while before we commence the construction of conscious artificial intelligence.

[32] A popular quote from Star Trek: "Superior ability breeds superior ambition".

References

Berman, R. & Piller, M. [Creators]. Star Trek: Deep Space Nine. Season 3 episode 2, “The Search: Part 2”.

Bostrom, N. Superintelligence: Paths, Dangers, Strategies [1st Edition]. (2014).

Chalmers, D. The Singularity: A Philosophical Analysis. (2010).

Dostoyevsky, F. Crime and Punishment. (2000). Woodsworth Editions Limited.

Dworkin, Gerald, "Paternalism", *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/paternalism/>.

Garland, A. [Director]. Ex Machina. Universal Studios. 2014. Film.

Goertzel, B. Should Humanity Build a Global AI Nanny to Delay the Singularity Until It’s Better Understood? *Journal of Consciousness Studies*, 19, No. 1-2, pp. 96-111. (2013)

Omohundro, S.M. The Basic AI Drives. *Self-Aware Systems, Palo Alto, California*. (2008).

Roddenberry, G. [Creator]. Star Trek: The Original Series. Season 1, episode 23, “Space Seed”.

Sullins, J.P. When is a Robot a Moral Agent? (2006)

Weir, P. [Director]. Dead Poets Society. Buena Vista Pictures Distribution. 1989. Film.