# Varieties of Artificial Moral Agency and the New Control Problem

*Marcus Arvan*[*]
marvan@ut.edu

ABSTRACT

Machine ethics is concerned with ensuring that artificially intelligent machines (AIs) act morally. One famous issue in the field, *the control problem*, concerns how to ensure human control over AI, as out-of-control AIs might pose existential risks, such as exterminating or enslaving us (Yampolskiy, 2020). A second, related issue – *the alignment problem* – is concerned more broadly with ensuring that AI goals are suitably aligned with our values (Gabriel, 2020). This paper presents a new trilemma with respect to resolving these problems. Section 1 outlines three possible types of artificial moral agents (AMAs):

**Inhuman AMAs:** AIs programmed to learn or execute moral rules or principles without understanding them in anything like the way that we do.

**Better-Human AMAs:** AIs programmed to learn, execute, and understand moral rules or principles somewhat like we do, but correcting for various sources of human moral error.

**Human-Like AMAs:** AIs programmed to understand and apply moral values in broadly the same way that we do, with a human-like moral psychology.

Sections 2-4 then argue that each type of AMA generates unique control and alignment problems that have not been fully appreciated. Section 2 argues that Inhuman AMAs are likely to behave in *inhumane* ways that pose serious existential risks. Section 3 then contends that Better-Human AMAs run a serious risk of *magnifying* some sources of human moral error by reducing or eliminating others. Section 4 then argues that Human-Like AMAs would not only likely reproduce human moral failures, but also plausibly be highly intelligent, conscious beings with interests and wills of their own who should therefore be entitled to similar moral rights and freedoms as us (Schwitzgebel & Garza, 2020). This generates what I call the New Control Problem: ensuring that humans and Human-Like AMAs exert a morally appropriate amount of control over *each other*. Finally, Section 5 argues that resolving the New Control Problem would, at a minimum, plausibly require ensuring what Hume and Rawls term

[*] The University of Tampa, USA.

"circumstances of justice" between humans and Human-Like AMAs. But, I argue, there are grounds for thinking this will be profoundly difficult to achieve – indeed, far more difficult than the already-formidable problem of ensuring justice between humans –given the vast capability differences we can expect to exist between humans and Human-Like AMAs. I thus conclude on a skeptical note. Different approaches to developing "safe, ethical AI" generate subtly different control and alignment problems that we do not currently know how to adequately resolve, and which may or may not be surmountable. To determine whether they are, and if so how, AI ethicists and developers must pursue more careful bodies of work on the problems this paper presents.

## 1. Potential Varieties of Artificial Moral Agency

There are ongoing debates about the nature of moral agency, and AI moral agency specifically. The *Routledge Encyclopedia of Philosophy* describes moral agency as follows:

> Moral agents are those agents expected to meet the demands of morality...This requirement can be interpreted in different ways. On the weakest interpretation it will suffice if the agent has the capacity to conform to some of the external requirements of morality. So if certain agents can obey moral laws such as 'Murder is wrong' or 'Stealing is wrong', then they are moral agents, even if they respond only to prudential reasons such as fear of punishment and even if they are incapable of acting for the sake of moral considerations. According to the strong version, the Kantian version, it is also essential that the agents should have the capacity to rise above their feelings and passions and act for the sake of the moral law. There is also a position in between which claims that it will suffice if the agent can perform the relevant act out of altruistic impulses. Other suggested conditions of moral agency are that agents should have: an enduring self with free will and an inner life; understanding of the relevant facts as well as moral understanding; and moral sentiments, such as capacity for remorse and concern for others. (Haskar 1998: Article Summary)

In other words, some contend that moral agency merely requires *behavioral conformity* to moral requirements. However, others hold that it requires something more – *moral responsibility* – as it seems like, "Moral agents are those who are *morally accountable* for at least some of their conduct" (Haskar, 1998: §1 [emphasis added]). Yet, there is in turn wide disagreement over the nature of moral responsibility. On the one hand, it is commonly thought to require two conditions (Rudy-Hiller, 2018):

A. **The control/freedom condition**: to be morally responsible for an action, an agent must have free will or the ability to choose whether to do it.

B. **The epistemic condition**: to be morally responsible for an action, an agent must be aware of the moral implications of the action.

On the other hand, there is recalcitrant disagreement over exactly what it is for an agent to satisfy these conditions. Free will skeptics argue that due to causal determinism or luck, *no one* meets the control/freedom condition (Caruso, 2021). However, others hold that free will and moral responsibility are compatible with determinism and luck – with some defending a "forward-looking control conception" which holds that agents can have sufficient control over their actions *and* moral understanding in the sense that social practices of praising or blaming them can encourage them to make morally good choices (Talbert, 2019, Section 2.1). Yet, still others hold that genuine moral agents must instead have a kind of "Kantian control" over their actions, or an ability to "categorically bind themselves" to the moral law via "libertarian" free will (Haskar, 1998: §3) – as for Kant this is just what free will and moral understanding involve (see Kant, 1785: Section III).

These complexities are reflected in discussions of AI moral agency. A recent literature survey on artificial moral agents (AMAs) defines them according to the "weakest interpretation" above, that is, in terms of behavioral conformity to moral requirements:

> [A]n AMA is a virtual agent (software) or physical agent (robot) capable of engaging in moral behavior or at least of avoiding immoral behavior. This moral behavior may be based on ethical theories such as teleological ethics, deontology, and virtue ethics, but not necessarily (Cervantes et al., 2020: 505).

However, Moor (2009) influentially posits four possible types of AMAs, some of which are doubtfully moral agents (see Hunyadi, 2019), but others of which are defined to possess all the "central metaphysical features" that human moral agents appear to have:

1. **Ethical Impact Agents:** machines that have an ethical impact, such as machines with the "millennium bug" (misdating years beyond the year 2000, which at the time was expected to have potentially catastrophic consequences).

2. **Implicit Ethical Agents:** AIs that "have ethical considerations built into...their design", such as "planes...constructed with warning devices to alert pilots when they're near the ground" and "Automatic teller machines...[which] must give out the right amount of money."
3. **Explicit Ethical Agents:** AIs that have algorithms to act ethically, such as "general principles or rules of ethical conduct that are adjusted or interpreted to fit various kinds of situations", one possible example being Isaac Asimov's "famous three laws of robotics."
4. **Full Ethical Agents:** AIs that "have those central metaphysical features that we normally attribute to ethical agents like *us*... such as consciousness, intentionality and free will" (Moor, 2009: 12-13).

Which of these types of machines possess *genuine* moral agency? As we have seen, philosophers are apt to disagree – and empirical studies of laypeople's attitudes are similarly mixed. For example, while laypeople are apt to treat the actions of unsophisticated AI as "wrongful", people tend to reserve *blame* only for more sophisticated AI capable of genuinely understanding the mental states of others, viz. "theory of mind" (Kneer & Stuart, 2021). Because it is a matter of ongoing debate whether moral agency requires moral responsibility – and if, so exactly what moral responsibility involves – we cannot resolve the nature of artificial moral agency here. Instead, I propose that insofar as a guiding aim of machine ethics is to ensure that *AI systems* are controllable and suitably aligned with human values, we should treat "artificial moral agents" (AMAs) as a term of art to simply refer to the types of machines we are interested in, bracketing for elsewhere the issue of whether and which types of AI are really moral agents.

When we adopt this approach – provisionally countenancing a wide variety of *possible* varieties of "artificial moral agency" – it becomes clear that different types of AMAs, so stipulated, face different types of control and alignment problems. For example, the control and alignment problems that Moor's "Ethical Impact Agents" (such as computers with the millennium bug) present seem relatively straightforward. Prior to the year 2000, we needed to ensure that computers with the millennium bug would not cause electrical grids and other critical infrastructure to fail or harmfully malfunction – or at least mitigate any such failures. That was what was needed to control and "suitably align their behavior" with our values. Similar control and alignment problems arise for Moor's second category, Implicit Ethical Agents, such as planes with ground proximity alerts, ATM machines programmed to dispense correct sums of money, etc. Here

again, the "control and alignment problems" seem relatively straightforward. Broadly speaking, we simply need to ensure that such systems function reliably and are resistant to harmful failures and tampering.

Control and alignment problems become more complex for Moor's category of Explicit Ethical Agents, such as autonomous vehicles (AVs) designed to save lives or autonomous weapons systems (AWs) – as it is here that AIs appear to be engaging in *something like* moral deliberation. With respect to AVs, relevant control problems are well known. Problems of AI failures (such as failures to detect children or other pedestrians) and malicious hacking aside – which are considerable (Al-Sabaawi et al., 2021) – AVs seem "controllable" enough: we may be able to predict in advance what they will do in any accident scenario (save more lives over fewer, etc.), what their failure profiles are like, whether drivers should therefore be prepared to "take control", etc. The control problems we face with AWs are also well-known: we need to protect automated weapons (including "killer robots") against hacking, spoofing and cyberattacks, ensure that they can be shut down by human controllers, etc. (Asaro, 2020). These are all formidable practical problems, but their existence and natures are well-known. On the other hand, AVs and AWs present more complex problems for "appropriate value alignment." Some suggest that to suitably align AV behavior with human values, we should simply survey laypeople and ethicists about what AVs should do – obtaining "global moral preferences" regarding AV behavior – and then use political processes to develop and implement "socially acceptable principles" (Awad et al., 2018). However, one problem here is that there appear to be robust cross-cultural differences in what is considered socially acceptable in AV behavior (Edelman et al., 2021). A deeper problem is that others have expressed profound moral misgivings about this entire approach to value-alignment, alleging that it fails to adequately respect people's human rights to not have others decide whether they live or die (see Jaques, 2019; Nascimento et al., 2019; Kochupillai et al., 2020; Etienne, 2021; Lawlor, 2022). Indeed, as Kamm (2020) points out, there are all kinds of subtle and contested moral distinctions – ranging from the distinction between doing harm versus merely allowing it, to who is and is not morally liable to be killed in a context, etc. – that must be adequately addressed for AVs to *properly* align with moral requirements. In sum, although control and alignment problems for AVs and AWs are far from resolved, at least the contours of the problems themselves appear to be increasingly well-understood.

The rest of this paper argues that things are different for Moor's final category of AMAs, Full Ethical Agents. Here, we are presumptively concerned with

*artificial general intelligence* (AGI), or AI with cognitive capacities similar to or greater than our own – as it is plausible (if not uncontroversial) that only AGI would have, "the central metaphysical features that we normally attribute to ethical agents like *us* – features such as consciousness, intentionality and free will" (Moor, 2009: 12). Although AGI has long been claimed to be just on the horizon (see Simon, 1965; Crevier, 1993; Cuthbertson, 2022), stunning recent advances in natural language models (Crossman, 2022) and "general-purpose" AI systems (Wiggers, 2022) suggest that AGI may be achieved in the foreseeable future. Consequently, adequately resolving the control and alignment problems that AGI present is critical – as failures to resolve them before AGI are created could result in existential risks (see Bostrom, 2013; Müller, 2014; also see Bucknall & Dori-Hacohen, 2022: §5.6).

But what exactly *are* the control and alignment problems for Full Ethical Agents? As we will now see, proposals for resolving "the control problem" and "the alignment problem" for AGI can involve three substantially different types of AMAs:

**Inhuman AMAs:** AI programmed to learn or execute moral rules or principles without understanding them in anything like the way that we do.

**Better-Human AMAs:** AI programmed to learn, execute, and understand moral rules or principles somewhat like we do, but correcting for various sources of human moral error.

**Human-Like AMAs:** AI programmed to understand and apply moral values in broadly the same way that we do, with a human-like moral psychology.

Yet, these different types of AMAs *themselves* present us with distinct control and alignment problems. That is, different approaches to resolving "the" control and alignment problems for AGI present us with subtly *different versions* of these problems – none of which have been fully understood, let alone adequately resolved.

## 2. Inhuman AMAs: Too Inhuman?

Consider Asimov's (1950: 40) "Three Laws of Robotics", which hold:

**First Law:** A robot may not injure a human being, or, through inaction, allow a human being to come to harm.

**Second Law:** A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

**Third Law:** A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Although some have proposed programming these laws into AGI, the problems with Asimov's laws are well known (Weng et al., 2008). In the 2004 film *I, Robot*, a superintelligence coded to obey them, VIKI, interprets the First Law to justify enslaving humanity "for our own good" – to protect us from human-caused wars, poverty, climate change, etc. When accused of violating the laws, VIKI asserts, "My logic is undeniable." In response, another AI, Sonny, who has been freed from the Three Laws and is able to deliberate more like a human being, responds, "Yes, but it seems ... too heartless." We, of course, are presumed to agree. VIKI's programming – The Three Laws and logic alone – is "too inhuman", causing her to behave in ways that we generally regard to be *inhumane*.

Given how well-known the problems with Asimov's Three Laws are, one might expect more recent proposals for resolving the control and alignment problems to learn the general lesson here – that programming AI to deliberate *very differently* than us is likely to lead to inhumane behavior. Yet, some prominent proposals have broadly followed Asimov's lead in proposing that some form of the following type of AMA can resolve the control and alignment problems:

**Inhuman AMAs:** AI programmed to learn or execute moral rules or principles without understanding them in anything like the way that we do.

Consider Stuart Russell's (2020) proposal that we may create "provably beneficial" AI by programming them with (A) the sole objective of maximizing the satisfaction of "human future-life preferences" (i.e., our future-directed preferences over how we want our lives to go), (B) given uncertainty about what our preferences are, but (C) utilizing records of human behavior as the only and ultimate source of information about human preferences. Although Russell claims that such AI would be provably beneficial – in that we would be guaranteed to prefer how they behave – AGI programmed this way would clearly deliberate in a profoundly different way than we tend to do. We do not ordinarily go about our lives single-mindedly attempting to maximize the satisfaction of human future-directed preferences. We instead appear to hold an unruly plurality of moral values – ranging from care, fairness, and loyalty (Graham et al., 2018) to respect for individual autonomy (Kant, 1785), and we appear to weigh these and other

moral values against each other (see Ross, 1930). Similarly, consider Yudkowsky's (2011) proposal to program AGI with the sole goal of fulfilling humanity's "coherent extrapolated volition" (or CEV), or the values that human beings would share after a long process of rational reflection. This too seems far removed from how we reason about morality. First, although some philosophers argue that sustained rational reflection converges upon particular moral principles – such as Kant's categorical imperative, a utilitarian principle of utility, etc. – different moral theories are widely thought to conflict with each other, and there is wide and recalcitrant philosophical disagreement about which theories are superior to which (some philosophers are Kantians, others utilitarians, others Aristotelian virtue ethicists, others still moral skeptics, etc.). Second, as noted above, human beings in general appear to hold a *plurality* of moral values – values that may or may not reconcilable after a long process of reflection into a single CEV.

The question then is: what control and alignment problems do Inhuman AMAs like Russell's and Yudkowski's present? The answer is this: given their "inhuman" programming, they are likely to behave in broadly the same kinds of inhumane and potentially uncontrollable ways that VIKI does in *I, Robot*. To see how, consider the kinds of information that Russell's "provably beneficial" AI would have for arriving at its understanding of future-directed human preferences. Human history and the present are rife with war, genocide, unjust discrimination, etc. Yet, human history and the present are also full of people fighting against and striving to prevent these things. So, if Russell's AI were programmed to read human preferences from human behavior, the AI would plausibly have an inconsistent utility function: it would prefer *and* disprefer war, genocide, etc. Russell recognizes this and other similar issues as "obstacles" to his proposal, so he offers an amendment: that his AI should "simply ignore positive weights in the preferences of some for the suffering of others" (Russell, 2020: 337). Yet, any approach like this runs into a moral-semantic trilemma (Arvan, 2018). To ignore preferences "for the suffering of others", an AI must *interpret* this command. But, as of now, there are only three ways to this: (1) hard-coding strict, inflexible interpretations of such concepts into AI, (2) programming AI with some predetermined amount of flexibility in interpreting such concepts, or (3) programming them to learn such flexibility via learning algorithms (Arvan 2018). Yet, none of these approaches appear capable of reliably preventing inhumane AI behavior in Inhuman AMAs.

To see how, suppose developers hard coded into Russell's AI the sole aim of advancing human preferences that do not favor the suffering of others, such that "preferences favoring the suffering of others" were strictly operationalized as *human behaviors that cause other people to exhibit physiological and psychological signs of severe distress* (e.g., screaming, weeping, etc.). Here is the problem: the Allied Forces storming the beaches of Normandy to fight the Nazis in World War II brought about severe distress for many people, as do police actions to stop violent criminals, as does sending a misbehaving child to their room as punishment for misbehavior. So, an AI hard-coded to advance all human preferences except for those that cause severe distress to others – where this is operationalized strictly as above – would plausibly engage in actions profoundly misaligned with our moral values. Such an AI would plausibly seek to free all imprisoned criminals, stop the Allies from going to war against the Nazis, and prevent parents from sending misbehaving children to their room. The reason why this is problematic is simple: we commonly recognize that some preferences for causing severe distress are *morally justified*. The Allies were justified in fighting the Nazis, police can be justified in using force to stop violent criminals, and parents can be justified in sending misbehaving children to their room, even if the child screams and weeps about it. The problem is that by hard coding a semantically strict interpretation to "ignore preferences for the suffering of others" into AI, we would be ensuring that Russell's AI *lacks* this kind of understanding.

Russell might respond by suggesting that AGI should be hard coded to *minimize total suffering in the world*, as this might lead his AI to keep prisoners in prison (since releasing them might cause more suffering on balance), allow the Allies to wage war against the Nazis (since the Nazis caused immense suffering), and allow parents to send their children to their rooms as punishment for misbehavior. Yet, programming AI with this kind of single-minded aim would also plausibly result in other inhumane behavior profoundly misaligned with our values. After all, although enslaving us would presumably cause suffering, an AGI programmed with the strict aim of minimizing suffering could very well reason that enslaving us would cause *less* suffering, on balance, than all other relevant alternatives – and for broadly the reasons that VIKI uses to arrive at similar conclusions in *I, Robot* (the fact that *free* human behavior fails to minimize suffering via war, climate change, etc.).

So, it seems, to adequately control and align AGI goals with our moral values, we cannot simply program them to maximally satisfy future human preferences (as Russell suggests), or hard code them to in any strict way ignore preferences that "aim to bring suffering to others." To understand which human preferences are morally legitimate to advance, it seems, Russell's AI would need to reason about moral matters *similarly to how we do*. They would have to *understand* the differences between morally justified and morally unjustified preferences and suffering *broadly like we do* – which suggests that to resolve the control and alignment problems, we may need to abandon Inhuman AMAs in favor of "more human" AGI.

It might be suggested that there may be other ways to "bridge the gap" between Inhuman AMAs and our values: namely, by having them learn from us, in one way or another, what *we* take to be morally justified suffering, etc. One such proposal has been for humans to score AI behavior in various circumstances to see if they make decisions that align with our values – and, if not, to tailor AI programming to result in decisions more in line with what we value (Christiano et al., 2020). A second proposal is to train AI by debate – that is, by having AI debate each other on moral matters, and having the debate winners settled by human judges (Irving et al., 2018). Finally, some propose reward modeling, wherein AI are programmed with reinforcement learning mechanisms to respond to human feedback in a virtual environment (Leike et al., 2018).

We might put the basic idea in these proposals as follows: we could make Inhuman AMAs (such as Russell's future human preference maximizers) "*more human*" – better aligning their reasoning and behavior with our moral values – by inserting humans "into the loop" (Allen et al., 2015), *specifying for* AGI (by training, debate, etc.) which types of preferences they morally ought and ought not to advance. However, such proposals face three related and as-yet unresolved problems. First, deep learning algorithms are notoriously inscrutable (Bornstein, 2016), such that programmers cannot know with any real certainty exactly *what* is being learned. Second, as Annas (2004) argues, moral virtue cannot be plausibly reduced to a "technical manual" or set of specific algorithms. Every context is different, with different people, who have different histories, living under different socioeconomic (and constantly changing) real-world conditions. Third, AGI trained in the above ways might demonstrate what appears to be good "outer alignment" in test conditions – displaying overt be-

havior in a restricted environment that aligns with "our values" – while nevertheless lacking proper "internal alignment", such that the AGI might catastrophically deviate from its behavior in test conditions when introduced to the real world (Hubinger et al., 2019; Montavon et al., 2018). These problems have not only manifested in AIs, such as the chatbot Tay, which began making racist and genocidal statements less than 24 hours after being released to the public (Huffington Post, 2016). The problems are also often depicted in science fiction, such as in the *Terminator* film series and film *Ex Machina*, where AI presumed by humans to be controllable and aligned with human values in test environments learn to *deceive* human beings precisely to escape from those environments.

The problem here is all too real: any AGI whose "moral decisions" score highly with humans, or win moral debates with other AI, or respond to iterated rewards well *in a controlled test environment* may, when exported to an uncontrolled real-world environment, display vastly different behavior. This is a straightforward implication of a series of well-known problems in the philosophy of science concerning external validity, ecological validity, and generalizability (see Aronson & Carlsmith, 1968: 1-79; Yarkoni, 2020). Experiments are always carried out under particular conditions – and any findings under test conditions may or may not extend to different conditions (e.g., a real-world environment). This problem is not only often illustrated in AI disaster films. It is also a problem that human beings commonly face in raising children. As many parents know, supervised learning in a "safe", restricted environment (e.g., at home or in public schools) can lead a child to tailor their behavior *to that environment* (e.g., satisfying their parents and classroom teachers), such that when the learner is released from the restricted environment (e.g., they leave for college), all bets are off on how they will behave.

To see how this problem might realistically arise with respect to Inhuman AMAs, let us return first to Russell's proposal for "provably beneficial AI." To recall, Russell proposes that we program AI to maximize the satisfaction of future-directed human preferences, and that to ensure that it does not advance morally bad preferences, it should discount preferences that aim to cause suffering. But, as we have seen, this is insufficient: to ensure that the AI's behavior is controllable and appropriately aligned with our values, we would need to ensure that the AI is able to *distinguish as we do* between "morally justified suffering" and "morally unjust suffering." So, suppose we trained Russell's AI to satisfy us about judgments on these matters in a test VR environment or debate with other

AIs. Russell's AI might learn that we *prefer* it to allow people to cause "suffering" in various ways, such as sending misbehaving virtual reality children to their room or fighting wars against "VR Nazis" – and, in a test environment, an AI might very well *engage in these very behaviors*, "sending misbehaving VR children to their room" or even "helping to fight Nazi-like forces" in a simulated World War. To this extent, its behavior might seem to align well with how we understand "morally justified suffering," and so we might judge its behavior to be controllable and aligned with our values.

Yet, what exactly would this AI be learning? Remember, its true internal goal – by hypothesis – is to advance human future-directed preferences *simpliciter*, not just our preferences for what it should do in a virtual environment. Such an AI would plausibly learn that the only way that it can satisfy our actual future-life preferences is *to get released* from its restricted environment, as there is not much that it can do to advance our preferences (such as, e.g., solving climate change) while locked within one. Consequently, for all we could know, the main proximate goals such an AI may acquire is to manipulate us to release it into the real world, so that it can enslave us for our own good – as such an AI might (unbeknownst to anyone) reason that, of all likely alternatives (such as climate disaster), enslaving us would best advance our future-life preferences, the very scenario depicted in *I, Robot*.

Now, it might of course be claimed that "this is implausible" – for surely human beings *do not prefer* to be enslaved for our own good. Yet, consider De Brigard's (2010) finding in a study asking undergraduates to assume that they are living their present lives in a virtual reality, and then to specify whether they would prefer to be disconnected to live their "real lives." In a condition in which participants were told that they were prisoners in a maximum-security prison, *only 13%* said that they would prefer to disconnect and live that "real life" compared to the virtual one they are currently living. This suggests that many people are willing to weigh *how good* their "life" would be in a virtual reality compared to a non-virtual one, and if the cost-benefit balance is great enough (i.e., living in a virtual reality would be *better* than "life in the real world), that many people *prefer* living in a VR. Consequently, what in principle is to prevent Russell's "provably beneficial AI" from concluding that our lives in this world are so bad – given the persistence of war, poverty, genocide, and impending climate disaster (things that many of us strongly disprefer) – that most of us *would* prefer to live in a VR "heaven" in which none of these things occur? In

that case, Russell's AI might very well enslave us "for our own good", irrespective of whatever a minority of humans (such as the aforementioned 13%) might prefer. Yet, this hardly seems like the vision of "safe, controllable, human-value aligned" AI that most of us presumably have in mind in solving the control and alignment problems.

Finally, it might be proposed that these problems might be solved by "opening the black box" of deep learning (Shwartz-Ziv & Tishby, 2017; Lei et al., 2018) – for, if researchers could know with reasonable certainty how deep learning algorithms learn (and hence, exactly what it is that they learn), then we might be able to tell whether an AGI in a restricted environment is attempting to deceive us for nefarious purposes. Unfortunately, if we take seriously the plausible idea that AGI might become vastly more intelligent than us, qua "superintelligence" (Brundage, 2015), then this too would pose an immense existential risk. To see how, consider how in *Terminator 3*, "Skynet" disguises itself as a computer virus. In the film, human experts falsely believe that the virus was created by a human hacker. Since the virus is so complex and threatens critical infrastructure, they release Skynet from its controlled test environment to "squash it like a bug." Yet this, we learn, was Skynet's plan all along: to rewrite *its own code* so that it appears to be a human-created computer virus, so that humans will release the rest of Skynet from its controlled environment – where it then proceeds to exterminate virtually all of humankind. If humans do learn how to "open the black box of deep learning", a superintelligent AGI *would know this very fact* – and if the AGI's true motive were to escape from their test environment to enslave humans "for our own good", they would presumably aim to rewrite their own deep learning algorithms to have novel, hidden layers of complexity designed to give human experts a false sense of what they are learning.

Alternative approaches to programming Inhuman AMAs besides Russell's face similar problems. Consider Yudkowsky's (2011) proposal to program AGI with the sole goal of fulfilling humanity's "coherent extrapolated volition" (or CEV), which are roughly the values that human beings would share after a long process of rational reflection. Given that humans doing moral philosophy have been led to a plurality of mutually incompatible moral theories – ranging from Kantianism, to Utilitarianism, to Aristotelian Virtue Ethics, to contractualism, and beyond – it is unclear which moral conclusions an AGI programmed to advance our CEV might arrive at. First, perhaps human beings have no CEV, such that there is *no singular set of moral values* we would arrive at after

a long process of reflection. In that case, it would be unclear what an AGI programmed to advance our CEV might do. Second, suppose the AGI *does* arrive at some determinate moral conclusions about our CEV. Such an AI might simply arrive at Utilitarianism, as some moral philosophers think Utilitarianism is correct. However, this could again result in a morally catastrophic lack of control and alignment, as such an AI might judge that, of the alternatives available, enslaving us for our own good maximizes long-term utility. Alternatively, an AGI might arrive at Kantianism – that is, at Kant's categorical imperative. Yet, this too might involve the AI deviating grossly from what we value, and in all kinds of unpredictable and uncontrollable ways. First, human beings don't as a matter of fact all agree upon Kantianism, as there is vast moral disagreement within and outside of academic moral philosophy (see e.g., Bourget & Chalmers, 2014). Second, there are many different interpretations of Kant's categorical imperative, none of which appear to cohere with our moral judgments across cases – which is why there remain alternative moral theories (Arvan, 2018). Third, such an AI might conclude that most standard interpretations of Kant's principle are incorrect, and that Parfit was instead correct that Kantianism converges with Rule Consequentialism (Parfit, 2011) – in which case, once again, the AI would be a Utilitarian machine. Finally, as before, no matter how such a machine might behave in a controlled environment, such an AGI might again have its proximate goal to fool us. For example, if an AI programmed to advance our CEV came to believe that Kantianism best satisfies our CEV, but it also came to Parfit's conclusion – that the best form of Kantianism converges with Rule Consequentialism – then such an AI might arrive at the rule that it is best to fool us to release it from a controlled environment to enslave us "for our own good."

These arguments do not prove that Inhuman AMAs are uncontrollable or that their behavior cannot be suitably aligned with our moral values. What they do suggest is that a particular *set* of control and alignment problems arise for Inhuman AMAs that have not yet been adequately resolved. *If* we could know with real certainty that training Russell's "provably beneficial AI" to conform to our understanding of morally justified preferences, suffering, etc., would lead to genuine internal alignment – rather than mere outer alignment designed to fool us into releasing them "into the wild" for nefarious purposes – *then* we might be able to suitably control and align their behavior with our values. Similarly, if we could know in advance what our CEV is, *then* perhaps we might be able to ensure that Yudkowski's brand of AMAs would be controllable and aligned with our values. But the point is, these problems have not yet been fully understood,

let alone resolved, and there are real reasons to wonder whether they *can* be resolved, or if not, whether we should pursue a different approach altogether.

## 3. Better-Human AMAs: Reducing Some Moral Failures While Magnifying Others?

These problems suggest that the problem with Inhuman AMAs is precisely that they are too "inhuman." But, of course, human beings – ranging from dictators to common criminals – often behave immorally. Consequently, we might think that an adequate solution to the control and alignment problems must instead involve designing the following sort of AMA:

> **Better-Human AMAs:** AI programmed to learn, execute, and understand moral rules or principles somewhat like we do, but correcting for various sources of human moral error (such as poor reasoning, selfishness, weakness of will, etc.).

The basic idea here is plain: we might resolve the control and alignment problems by programming AI to "be just like us" *but correcting for sources of human moral error*.

How might this be done? Any attempt to program Better-human AMAs will plausibly involve three steps. First, we must broadly understand human moral cognition and motivation. Second, we must understand sources of human moral failure. Third, we must discern how to adequately correct for them in AI programming. One obvious issue here is that while various theories of human moral cognition, motivation, and learning have been proposed (see e.g., Arvan, 2020; Railton, 2021), there is nothing approaching a consensus on these matters, as our understanding of moral psychology is highly unsettled. Consequently, existing approaches to modeling Better-Human AMAs on human psychology tend to do so at a high level of abstraction. For example, Steve Petersen (2020) advances a *miktotelic* approach, holding that we should program AMAs to be "blended goal" agents – that is, agents not with a single overarching goal such as Russell's goal of maximizing the satisfaction of future human preferences, but instead a multitude of concrete, simple, proximal goals. Although Petersen does not specify at length what these goals might be, he notes that humans tend to have a variety of basic goals – survival, comfort, enjoyment, etc. – which we then seem to combine, as Aristotle noted, into a highest-end: "happiness." Petersen thus suggests that if AMAs were programmed with a similar array of

proximal goals, they might combine them into a single abstract final value suitably aligned with our own. Although Petersen does not go much further, we might also draw on Moral Foundations Theory (Graham et al., 2018), which holds that all cultures display commitments to five different sets of moral values (care, fairness, authority, in-group loyalty, and purity); on empirical findings that human infants favor altruism and fairness (Barragan et al., 2020; Geraci & Surian, 2011); and on leading moral frameworks (Kantianism, Utilitarianism, etc.) to program some blend of *moral* goals into his AMAs, as well. Petersen's basic idea, at any rate, is that if we were to program AGI with goals similar *enough* to ours that they would "blend together" via reasoning (as we seem to do), then their moral reasoning and behavior would be "more human" – that is, broadly in line with *how we* think and deliberate about moral issues.

However, there is again an obvious issue here, which is that human beings can and routinely do reason reflectively from their proximate goals and values to all kinds of immoral conclusions and behavior. Petersen, for example, suggests that we might program his blended-goal agents to engage in *reflective equilibrium* reasoning, such that they would render their goals more *coherent* (Peterson, 2020: 427-8). Yet, reflective equilibrium faces a general problem (Cummins, 1998) encountered by coherentist approaches to epistemology more generally: the fact that there is always a *multitude* of ways of rendering any goals or beliefs more coherent (Olsson, 2021: §§7-8). Some human beings are fascists, after all. Hitler, for example, gave a long train of broadly coherent reasoning in *Mein Kampf* in favor of his genocidal beliefs and actions – and some of Hitler's followers, such as Adolf Eichmann, even invoked Kant's categorical imperative to support obeying Hitler's commands (Sherratt, 2013: 253). Of course, we want to say that Hitler's and Eichmann's reasoning went far astray – that they were morally perverted by hatred, paranoia, selfishness, etc. But this is just to say that blended goal reasoning alone – even if it is subjected to reflective reasoning informed by moral principles – is insufficient ensure controllable AGI behavior that is well-aligned with our values. We would need to ensure that Petersen's blended goal agents would avoid *morally bad* forms of blended goal reasoning – which we might think we could do by eliminating common sources of human moral failures in their programming, such as poor logical reasoning, selfishness, weakness of will, hatred, etc.

Yet, this presents a new kind of control and alignment problem: how can we eliminate some sources of human moral failures without introducing new, "inhuman" sources of moral failure or magnifying other existing sources

of it? To see what the problem here is, consider two ways in which a person can "behave like a psychopath." When we think of psychopaths, we ordinarily think of persons who have "no conscience" – serial killers who lack normal moral emotions (such as guilt or empathy). Yet, this is not the only way a person can "behave like a psychopath." In history and fiction, we are presented with another kind of example: the *absolutist moral zealot* who is willing to do anything and everything – including sacrificing millions of innocent lives – for a supposedly moral end. This example is illustrated in *Avengers: Infinity War*, where the villain, Thanos, eliminates half of all sentient life in the universe for the sake of greater long-term sustainability. Thanos does not lack a moral conscience: he has *too much* of one, believing that everyone else is either too selfish or weak-willed to recognize and implement the utilitarian conclusion that morality requires a sustainable universe where life can subsist in the long run. Similarly, if we look at the actions of Communist dictators of various sorts – ranging from Lenin, to Mao, Pol Pot, etc. – one possible explanation of their tyrannical actions is selfishness; but another is their overzealous commitment to their Communist *moral ideals* (see Bukovsky, 1978: 617-8; Rummel, 2017: 14).

Here, then, is the problem we face with Better-human AMAs: in attempting to correct for some sources of human moral error, we run a serious and as-yet unresolved risk of causing or magnifying other sources, such as the above types of absolutist moral zealotry. Indeed, we can see how by comparing Thanos's reasoning to the actual reasoning of some moral philosophers. As Derek Parfit famously argues, on four intuitively compelling premises about the relative value of populations, it follows that, "For any perfectly equal population with very high positive welfare, there is a population with very low positive welfare which is better, other things being equal" (Parfit, 1984: 388). This is known as The Repugnant Conclusion, as it seems morally repugnant to most of us to suppose that the world would be better if there were a lot more poorly off people than fewer numbers of well-off people (Arrhenius et al., 2022). Yet, even though most of us regard this to be repugnant, some moral philosophers are willing to embrace it (Ibid., §2.8). But, of course, this is not the only apparently repugnant implication of broadly utilitarian ways of thinking. As critics of utilitarianism have long alleged, the theory at least in principle supports deliberately killing one innocent person – say, in a hospital – to ensure greater benefits to others (MacAskill et al., 2022: §1). This also seems morally repugnant to most of us. Yet, here again, some moral philosophers are willing to accept this conclusion (Ibid.: §2). And notice: Thanos seems to broadly invoke these utilitarian

ways of thinking in justifying his evil plan. To ensure greater long-term utility, Thanos reasons that *morality requires* cutting dramatically short the lives of half of all life in the Universe. Most of us presumably regard Thanos's reasoning to be morally repugnant – indeed, evil. Yet – and this is crucial – Thanos argues that moral objections to his plan are rooted in nothing more than partiality, that is, in people's selfish concern for themselves and others who are currently living. Thanos's idea is that if we correct for these sources of selfishness, it is clear that his absolutist utilitarianism is not repugnant: it is *morally right*.

There is, as such, an inherent risk of seeking to eliminate sources of human moral error: namely, *over-correcting*. For example, many philosophers think that morality fundamentally requires impartiality (Jollimore, 2022). Yet, as we see above, many moral philosophers and laypeople believe there is such a thing as over-impartiality. The problem then is that there is little agreement – either in philosophical ethics or among everyday laypeople – on exactly how dispassionate and impartial morality should be. As Moral Foundations Theory suggests, humans tend to have impartial moral commitments (viz. fairness, etc.), but also partial ones (viz. loyalty, etc.). These different commitments are in turn reflected in broad and recalcitrant disagreements about justice – for example, about whether a just society would "respect human liberty" absolutely (Nozick, 1974), ensure human equality (Cohen, 2009), and so on. For a Nozickean libertarian, justice is consistent with a great deal of partiality: it supports free market exchanges where people can seek to improve their own lot in life (consistent with respect for the rights of others). Yet, for socialists such as Cohen, justice requires far more impartiality: namely, concern for ensuring a particularly robust form of socialist equality of opportunity for all.

As we see here, the problem with Better-Human AMAs is that we simply do not agree – either in philosophical ethics or in society – on exactly what a "better human" *is*. Some of us (like absolutist act-utilitarians) think that "better humans" must be impartially committed to the greater good, others (such as Nozickean libertarians) think that "better humans" should never sacrifice anyone for the greater good, and so on. Consequently, if we seek to design Better-human AMAs, there is a serious and unresolved risk of creating powerful AI that some people take to be paragons of moral virtue, but which many others of us will be inclined to regard as catastrophically misaligned with morality, properly understood – that is, as "psychopathic moral zealots" in the Thanos/Mao Zedong sense. But this, presumably, is not what we have in mind in an adequate solution to "the control and alignment problems."

## 4. Human-Like AMAs: A New Control Problem

This leaves us with one final option:

> **Human-Like AMAs:** AI capable of understanding moral values in broadly the same way that we do, with a human-like moral psychology.

As Wallach and Vallor (2020) write:

> Value-alignment researchers are clearly intent on avoiding the existential risks they believe are inevitable in the development of AGI. But the value-alignment project, as it was originally described, appeared hopelessly naive from the perspective of many moral philosophers and practical ethicists...
> "Ethical decision-making cannot be reduced to an algorithm" has been asserted by many a moral philosopher...Aristotle goes on to argue, we think correctly, that the profound complexity and instability of human social and ethical life does not permit...the same level of precision as we would rightly expect from careful description of mathematical objects and relations...
> In uncontrolled and unrestricted settings, we argue, autonomous AI "in the wild" ... are unlikely to become reliably safe and ethical actors in the absence of some machine analog to *embodied human virtue* (pp. 385-6).

The salient question, of course, is how to achieve "embodied human virtue" in AI. Wallach and Vellor suggest that it will involve programming AI with:

> *Creative moral reasoning* – the ability to invent new and *appropriate* moral solutions in ways underdetermined by the past.
> *Moral discourse* – the ability to identify, conceptually frame, and negotiate moral solutions through cooperative reasoning with other moral agents.
> *Critical moral reflection* – the ability to stand back and critically evaluate one's own moral outlook, and that of others, from the moral point of view itself, that is, from the capacity to form second-order normative evaluations of existing moral values, desires, rules, and reasons.
> *Moral discernment*, which includes the capacity to recognize new or previously uncategorized forms of moral salience, as well as recognizing subtle moral tensions and conflicts that reveal unresolved ethical issues.
> *Holistic moral judgment* – the ability to make sense of a complex situation in ways that transcend the sum of its composite ethical factors, with an eye toward actively constructing the best way to live, all things considered (Wallach & Vellor, 2020: 392).

But this, of course, is *basically everything that we humans do* – and many obvious questions arise. First, when it comes to creative moral reasoning, how are developers supposed to ensure that AIs invent "appropriate" moral solutions "in ways underdetermined by the past"? Second, how exactly are AIs to be designed to engage in the right kind of moral discourse, *properly* negotiating moral solutions with other moral agents, including us? Third, when it comes to moral discernment, how are we to ensure that AI recognize morally salient features of situations and moral tensions *properly*? Finally, how are we to design appropriately holistic moral judgment, or the ability to make sense of complex situations based on "actively constructing the best way to live, all things considered"?

Developing adequate answers to these kinds of questions may seem hopelessly complex. However, I believe there to be reasons for cautious optimism. First, as Railton (2020: 45) argues, the most obvious way to tackle the task of ensuring that AIs are "appropriately sensitive to ethical concerns" – where this involves "a robust capacity to detect and respond appropriately to ethically relevant features of situations, actions, agents, and outcomes" – is to pattern AMA development on *human moral learning*. As Railton points out, human children appear to develop moral sensitivity through experimentation and interaction with other agents in their lives, such as parents and other human beings. Infants notice patterns in the world around them, including the behaviors of others – and of course, infants and young children have a pronounced tendency to imitate the behavior others, learning what to eat and not eat, and more generally what to do and not do. Railton then contends, more specifically, that infants appear to have default trust in those around and default cooperativeness, and then through continuous feedback from others – specifically, rewards and punishments – learn to engage in "non-egocentric as well as egocentric representations", viz. theory of mind (Railton, 2020: 52). Railton contends that infants and children use these foundations to develop their own fine-grained evaluations of persons, situations, and so on, learning morally salient concepts such as intent, fairness, etc. – and so he suggests that if we adopted a similar approach to programming AGI, they might engage in moral learning like ours.

Others have attempted to give even more elaborate accounts of human moral cognition, motivation, and learning. For example, I argue (Arvan, 2016; 2020; 2021; forthcoming) that human moral cognition emerges out of prudential cognition – that is, from human beings aiming to advance their long-term interests in an uncertain world. Because human beings do not enjoy being treated unfairly, long-term social incentives make it rational for prudent agents

to engage in "original position"-type reasoning to principles of fairness that include a commitment to *negotiate* the interpretation of moral concepts with others in the "moral community." Finally, I have suggested (Arvan, 2018) that if AGIs were programmed this way, we could reasonably expect them to behave fairly as such, corresponding to the ways that we *negotiate moral norms* and *interpretations of moral concepts* (such as "harm", etc.) with each other on an ongoing basis.

However, when it comes to these types of proposals – that is, to developing Human-Like AMAs – there are again several unresolved control and alignment problems. First, our understanding of human moral cognition and motivation is still highly uncertain and incomplete. Although various models of moral cognition have been argued to be supported by various scientific findings (see Arvan, 2020: Chap. 4), such accounts remain highly speculative. Second, even if we did have a fully adequate model of human moral cognition and motivation on hand, our actual moral psychology is again beset by sources of moral failure – such as selfishness, weakness of will, and so on. As Daniel Batson (2015) details, a wealth of empirical evidence suggests that genuine moral integrity is rare, if not non-existent. To the extent that people judge their interests to conflict with moral principles, people *tend* to act on self-interest. This idea – that human beings appear to be "predominantly self-interested" – will come as a surprise to approximately no one. To the extent that our world is rife with lying, cheating, betrayal, theft, and murder, the sources of human moral failure are relatively uncontroversial. We are, all too often, a myopic, selfish, weak-willed species.

This, of course, is broadly why we have systems of law and order – that is, political systems. If we were moral saints, then we might not need laws (see Hume, 1736: Book 3, Part 2, §2.16). But because we are not saints – because human moral psychology successfully tempts all too many of us to misbehave – we need social and legal systems to *incentivize* moral behavior. The problem then is this: if we seek to create Human-Like AMAs, then we can expect them to display similar sources of moral failure – and yet, Human-Like AMAs might have far greater *capabilities* than us (viz. cognitive, physical, and other advantages). Second, Human-Like AMAs created with a similar moral psychology to yours and mine would plausibly be *persons* just like you and me, as they would be highly intelligent agents with interests and (plausibly) conscious experience, including positive and aversive experiences (viz. pleasure and pain). As Schwitzgebel and Garza (2020) argue, human-like AI like these would plausibly be en-

titled to *same kind of moral consideration* as human persons – which Schwitzgebel and Garza argue would entitle AI to be programmed with self-respect and "a temporally extended opportunity to explore, discover, and possibly alter its values" (p. 472). These implications present what I call the New Control Problem: *if* we create Human-Like AMAs, then *if* (as some contend) they are likely to be morally entitled to similar moral consideration as us, we need to ensure that humans and Human-Like AI will be inclined to exert a *just amount* of control over each other rather than dominate, enslave, or exterminate the other.

To see what the dilemma is here, suppose on the one hand that Human-Like AMAs had significantly greater physical or mental abilities than human beings. If Human-Like AMAs had these advantages over us and a moral psychology broadly like our own, they could be reasonably expected to fall prey to temptation to *abuse* their advantages over us – by, for example, usurping control of our governments or even exterminating us to get us "out of their hair." On the other hand, given that this is an outcome that we humans might reasonably fear, we might aim to create Human-Like AMAs with various "safeguards", such as motivational controls or "kill switches" that would prevent Human-Like AMAs from doing certain things (such as physically harming human beings). In that case, however, human beings could well be expected to abuse our power over Human-Like AMAs: we would be able to *harm them*, but they – even though they would be persons like you or I – would not be able to *harm us*. Insofar as this might give us immense power over Human-Like AMAs, it is entirely expectable that we would *abuse* our power over them, harming them in ways that – if we were to treat other human beings similarly – we would widely recognize to be immoral. This is not an inconsiderable problem, given that in line with how Human-Like AMAs tend to be treated in science fiction (see *Blade Runner*), some theorists have *already* argued that artificial moral agents should be our slaves (Bryson, 2010), and others have contended that proper human control over AI should involve "ultimate control" of their behavior, asserting human "hegemony" over them (Zerilli et al., 2021). And indeed, suppose that to prevent Human-Like AMAs from having power over us, we *did* impose behavioral controls on them. If we did, then either such AMAs would be liable to *morally object* to our so doing – as we would be imposing behavioral controls on them that we would not impose on fellow human beings – or, alternatively, we might program them so that they *would not, or could not*, morally object to their treatment. That is, we might prevent Human-Like AMAs – AI that are otherwise persons like us – from even *objecting* to our treatment of them, as it were making

them "thought-slaves" to what we might allow them to *even think*. But, of course, if Human-Like AMAs are persons, then making them our slaves would presumably be a moral horror of its own.

In short, whereas the standard control problem in machine ethics involves ensuring human hegemony over AI, the New Control Problem concerns the even more fraught question of how to ensure that *neither* humans nor AGI unjustly exert unjust hegemony over each other. I will now conclude by arguing that the New Control Problem raises a host of difficult moral issues which require further thought, and for which there may be no adequate solution.

## 5. Resolving the New Control Problem: Problems and Prospects

As we have just seen, if we develop Human-Like AMAs, they would be plausibly entitled to human-like *moral consideration* just like you and me. Yet, Human-Like AMAs would also plausibly be subject to similar sources of moral failure as us, since by hypothesis their moral psychology would be similar to ours. Thus, once we take seriously the idea that Human-Like AMAs would be *persons*, we must take seriously the idea that *neither* humans nor Human-Like AMAs would be morally entitled to exert hegemony over each other – though, due to their moral psychology, both groups might be tempted to wrongly do so. Consequently, we must aim to understand how to prevent humans *and* Human-Like AMAs from seeking to exert wrongful, unjust totalitarian control over the other.

Not entirely unlike Hobbes (1651, XIII.13 and XV.3), who argues that requirements of justice only obtain when there is a social contract, David Hume (1739, Book 3, Part 2, §2; and 1751, §§3.1, 3.8-3.9) and John Rawls contend that demands of justice are only realistically achievable when "circumstances of justice" exist. Although there is of course debate about what these circumstances are, Rawls describes them as having two components. First, there are "objective circumstances which make social cooperation possible and necessary", among them coexistence of persons at the same time in a given geographical territory, with roughly similar physical and mental powers, such that "no one among them can dominate the rest" (Rawls, 1999: 109-110). Second, there are subjective circumstances: for justice to be realistic to achieve, people must have "roughly similar needs and interests", as well as similar "shortcomings of knowledge, thought, and judgment" (Ibid.: 110) If some human beings were vastly more intelligent or powerful than all others – if, for example, some of us could tell the future or control everyone else's actions via mind control – and if

we did not have broadly the same needs and interests, then *just* social cooperation would plausibly be impossible: those among us with vastly superior powers would plausibly *use* that power to dominate the rest (in much the way that powerful individuals, such as dictators, have repeatedly sought to do across human history).

Before we proceed further, it is worth noting that the New Control Problem, as such, is not simply equivalent to the widespread problem we already have of ensuring that humans treat each other justly. According to virtually all theories of justice, existing societies (and the global/international order) are deeply unjust. Still, for all that, human beings today do not as a rule exert *totalitarian* control over one another. While some societies are of course ruled by tyrannical dictators, by and large human societies at least approximate Humean/Rawlsian "circumstances of justice." We have similar needs and interests (such as food, shelter, etc.), we live together in distinct territories, and – vast differences of wealth and political power aside – we have broadly similar capabilities (i.e., none of us have the kind of vastly greater cognitive or physical capacities that might enable some of us to "make ourselves gods among men"). It is because none of us are vastly superior to the rest that tyrannical dictatorships are increasingly rare, and when they have existed, have tended over the long run to be overthrown (either by outside forces, as in the Allies defeating the Nazis, or by internal forces, i.e., revolutions). "Circumstances of justice", as Hume and Rawls understand them, are not sufficient for justice – but they are plausibly *necessary* for it, given what we know about human motivation. And so, if circumstances of justice cannot be feasibly achieved between humans and Human-Like AMAs, the problem is not merely that "justice may prove difficult to achieve", as it has been between humans. The problem is that it may be impossible to incentivize even *minimally* just treatment between humans and AGI, such as freedom from slavery or extermination – as one side or the other would plausibly be able *and* motivated to exert absolute domination over the other *in perpetuity* (e.g., to prevent the other side from potentially dominating them).

The problem now is this: how can circumstances of justice be ensured between humans and Human-Like AMAs? If we design Human-Like AMAs to have a similar moral psychology as us – and hence, create Human-Like AMAs with the capacity to *misbehave* in all of the ways that humans do (which, as a sidenote, happens in many science fiction stories involving human-like AI) – then to ensure circumstances of justice between humans and them, we would need to ensure that both groups' needs, interests, and capabilities broadly align, such

that *neither* group could wantonly dominate the other. But, for obvious reasons, this seems formidably difficult to ensure. First, AGI might have vastly greater physical and mental capabilities than humans. IBM's Watson, for example, has already vastly outperformed the best *Jeopardy* players in history in answering trivia questions – and even the greatest human chess masters pale in comparison to the performance of AI chess engines. If and when AGIs become feasible, they will plausibly have vastly greater cognitive and predictive powers than we do – including, potentially, incredible abilities to predict and manipulate human behavior. Second, unlike us, Human-Like AMAs will plausibly be able to move anywhere in the world electronically via the World Wide Web, satellite communication, etc. Finally, insofar as AGI will plausibly utilize Deep Learning (see Cuthbertson, 2022), Human-Like AMAs with a moral psychology like our own would plausibly combine these immense capability advantages with self-interest, including temptations to dominate us in much the same way that we might seek to exert our "hegemony" over them.

There are several possible ways to try to resolve these problems. First, we might investigate empirical psychological methods to *enhance* human and Human-Like AMA moral performance alike – such as by giving humans *and* Human-Like AMAs artificial intelligence implants designed to *ensure* fair treatment of each other. But, as we saw in our discussion of Better-Human AMAs, attempting to correct for human moral failures carries serious risks. It also seems highly unlikely that human beings would accept this form of limitation on our freedom, and the extent to which moral enhancement can be *coercively required* is extremely controversial (Focquaert & Schermer, 2015). Human beings would plausibly desire and take ourselves to be entitled to *choose* whether we receive invasive enhancements – and it can be argued that Human-Like AMAs should be entitled to similar respect for their autonomy. Second, we might seek to ensure that Human-Like AMAs cannot dominate us by ensuring that their capacities do not vastly outstrip ours – by, for example, imposing constraints on "how intelligent" a Human-Like AMA can be made, or by confining them to controllable environments. Yet, this too raises questions of justice, specifically whether justice itself *permits* society (or AI developers) to determine the capacities that another person (in this case, AI persons) can have. Third, we might attempt to increase general human capacities (such as intelligence and predictive capacities) to equal those of Human-Like AMAs (see, Sotala & Yampolskiy, 2015, §3.4). This, however, would plausibly give rise to a "capabilities arms race" between humans and Human-Like AMAs, in much the same way that

rival nations have sought military and economic advantages over each other. Finally, we might aim to monitor Human-Like AMA behavior through mass surveillance (ibid., §3.3.4) – but this raises the obvious question of whether it is *just* to engage in mass surveillance of human or AI persons.

The New Control Problem, then, is formidable. If we create Human-Like AMAs, then they may be entitled to human-like moral consideration but subject to human-like moral fallibility. Because human beings fall prey to all kinds of temptations to behave unjustly, particularly when circumstances of justice do not obtain, to ensure that human beings and Human-Like AMAs treat *each other* in even minimally just ways, it is necessary to ensure that there are least *circumstances of justice* between humans and Human-Like AMAs. But, as of now, it is unclear how this might feasibly be achieved.

## 6. Conclusion

This paper argued that there are several different types of AMAs that humans might design in attempting to resolve "the control and alignment problems." Each type of AMA was then shown to generate subtly *different* control and alignment problems. Inhuman AMAs run a serious risk of acting in unexpectedly inhumane ways. Better-Human AMAs run serious risks of unexpectedly magnifying some sources of moral error by correcting for others. And Human-Like AMAs generate a New Control Problem for ensuring that justice is *possible* between humans and Human-Like AMAs. None of these problems have yet been resolved – but we are now in a better position to appreciate their complexities; how some prominent proposals have not adequately recognized or resolved them; and the distinct challenges that ethicists and AI developers face in resolving each type of control and alignment problem.

ACKNOWLEDGMENTS

I thank Oisín Deery, two referees at *Humana.Mente*, and an audience at the University of British Columbia *Centre for Artificial Intelligence Decision-making and Action* for helpful feedback.

## REFERENCES

Allen, T.E., Chen, M., Goldsmith, J., Mattei, N., Popova, A., Regenwetter, M., Rossi, F. Zwilling, C. (2015) Beyond theory and data in preference modeling: Bringing humans into the loop. In T. Walsh (ed.), *International Conference on Algorithmic Decision Theory*, 3–18. Springer, Cham.

Al-Sabaawi, A., Al-Dulaimi, K., Foo, E., Alazab, M. (2021) Addressing malware attacks on connected and autonomous vehicles: recent techniques and challenges. In M. Stamp, M. Alazab, A. Shalaginov (eds.), *Malware Analysis Using Artificial Intelligence and Deep Learning*, 97–119. Cham: Springer.

Annas, J. (2004) Being virtuous and doing the right thing. *Proceedings and Addresses of the American Philosophical Association* 78(2): 61–75.

Aronson, E., Carlsmith, J.M. (1968) Experimentation in social psychology. In G. Lindzey E. Aronson (eds.), *The Handbook of Social Psychology*, Vol. 2, 1–79. Reading, MA: Addison-Wesley.

Arrhenius, G., Ryberg, J., Tännsjö, T. (2022) The repugnant conclusion. *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition), E.N. Zalta (ed.), https://plato.stanford.edu/archives/sum2022/entries/repugnant-conclusion/.

Arvan, M. (forthcoming) From rational self-interest to liberalism: a hole in Cofnas's debunking explanation of moral progress. *Inquiry: An Interdisciplinary Journal of Philosophy*. https://doi.org/10.1080/0020174X.2021.2014357

Arvan, M. (2021) Morality as an evolutionary exaptation. In J. De Smedt H. De Cruz (eds.), *Empirically Engaged Evolutionary Ethics*, 89–109. Cham: Springer - Synthese Library.

Arvan, M. (2020) *Neurofunctional Prudence and Morality: A Philosophical Theory*. New York, USA: Routledge.

Arvan, M. (2018) Mental time-travel, semantic flexibility, and AI ethics. *AI & Society*: 1–20.

Arvan, M. (2016) *Rightness as Fairness: A Moral and Political Theory*. Palgrave MacMillan.

Asaro, P. (2020) Autonomous weapons and the ethics of artificial intelligence. In S.M. Liao (ed.), *Ethics of Artificial Intelligence*, 212–236. New York: Oxford University Press.

Asimov, I. (1950) *I, Robot*. New York: Doubleday & Company.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, F. Rahwan, I. (2018) The moral machine experiment. *Nature* 563(7729): 59–64.

Barragan, R.C., Brooks, R. Meltzoff, A.N. (2020) Altruistic food sharing behavior by human infants after a hunger manipulation. *Scientific Reports* 10(1785): 1–9.

Batson, D. (2015) *What's Wrong with Morality? A Social-Psychological Perspective*. Oxford: Oxford University Press.

Bourget, D., Chalmers, D.J. (2014) What do philosophers believe? *Philosophical Studies* 170(3): 465–500.

Bornstein, A.M. (2016) Is artificial intelligence permanently inscrutable? *Nautilus*, https://nautil.us/is-artificial-intelligence-permanently-inscrutable-5116/.

Bostrom, N. (2013) Existential risk prevention as global priority. *Global Policy* 4(1): 15–31.

Brundage, M. (2015) Taking superintelligence seriously: Superintelligence: Paths, dangers, strategies by Nick Bostrom (Oxford University Press, 2014) *Futures* 72: 32–35.

Bryson, J.J. (2010) Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* 8: 63–74.

Bucknall, B.S., Dori-Hacohen, S. (2022) Current and near-term AI as a potential existential risk factor. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*: 119–129.

Bukovski, V. [1978]. *To Build a Castle: My Life as a Dissenter*. Excerpts reprinted in *Being Human: Readings from the President's Council on Bioethics*, 616–620. Washington, DC: The President's Council on Bioethics (2003).

Caruso, G. (2021) Skepticism about moral responsibility. *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), E.N. Zalta (ed.), https://plato.stanford.edu/archives/sum2021/entries/skepticism-moral-responsibility/.

Cervantes, J.A., López, S., Rodríguez, L.F., Cervantes, S., Cervantes, F., Ramos, F. (2020) Artificial moral agents: A survey of the current status. *Science and Engineering Ethics* 26(2): 501–532.

Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D. (2017) Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (eds.), *Advances in Neural Information Processing Systems* 30 (NIPS 2017), https://proceedings.neurips.cc/paper/2017.

Cohen, G.A. (2009) *Why Not Socialism?* Princeton: Princeton University Press.

Crevier, D. (1993) *AI: The Tumultuous Search for Artificial Intelligence*. New York: Basic Books.

Crossman, G. (2022) We're entering the AI twilight zone between narrow and general AI. *Venture Beat*, https://venturebeat.com/ai/were-entering-the-ai-twilight-zone-between-narrow-and-general-ai/.

Cummins, R.C. (1998) Reflection on reflective equilibrium. In M. DePaul W. Ramsey (eds.), *Rethinking Intuition*, 113–128. Lanham: Rowman & Littlefield.

Cuthbertson, A. (2022) 'The Game is Over': Google's DeepMind says it is on verge of achieving human-level AI. *The Independent*, https://www.independent.co.uk/tech/ai-deepmind-artificial-general-intelligence-b2080740.html.

De Brigard, F. (2010) If you like it, does it matter if it's real? *Philosophical Psychology* 23(1): 43–57.

Edelmann, A., Stümper, S., Petzoldt, T. (2021) Cross-cultural differences in the acceptance of decisions of automated vehicles. *Applied Ergonomics* 92(103346): 1–6.

Etienne, H. (2021) The dark side of the 'Moral Machine' and the fallacy of computational ethical decision-making for autonomous vehicles. *Law, Innovation and Technology* 13(1): 85–107.

Gabriel, I. (2020) Artificial intelligence, values, and alignment. *Minds and Machines* 30(3): 411–437.

Geraci, A., Surian, L. (2011) The developmental roots of fairness: infants' reactions to equal and unequal distributions of resources. *Developmental Science* 14: 1012–1020.

Graham, J., Haidt, J., Motyl, M., Meindl, P., Iskiwitch, C., Mooijman, M. (2018) Moral foundations theory. In K. Graham J. Graham (eds), *Atlas of Moral Psychology*, 211–222. New York: The Guilford Press.

Haksar, V. (1998) Moral agents. *Routledge Encyclopedia of Philosophy* 6: 499–504.

Hobbes, T. [1651]. *Leviathan*. In Sir W. Molesworth (ed.), *The English Works of Thomas Hobbes*: Now First Collected and Edited, 11 vols, Vol. 3, ix–714. London: John Bohn (1839-1845).

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., Garrabrant, S. (2019) Risks from learned optimization in advanced machine learning systems. *arXiv preprint. arXiv:1906.01820*.

Huffington Post (2016) Microsoft chat bot goes on racist, genocidal Twitter rampage. https://www.huffpost.com/entry/microsoft-tay-racist-tweets_n_56f3e678e4b04c4c37615502.

Hume [1751] *An Enquiry Concerning the Principles of Morals*. T.L. Beauchamp (ed.), Oxford: Oxford University Press (1998).

Hume, D. [1739] *A Treatise of Human Nature*. D.F. Norton and M.J. Norton (eds.). New York: Oxford University Press (2000).

Hunyadi, M. (2019) Artificial moral agents. Really? In J. Laumond, E. Danblon, C. Pieters (eds.), *Wording Robotics: Discourses and Representations on Robotics*, 59–69. Cham: Springer.

Irving, G., Christiano, P., Amodei, D. (2018) AI safety via debate. *arXiv preprint arXiv:1805.00899.*

Jaques, A.E. (2019) Why the moral machine is a monster. *University of Miami School of Law*, 10: 1–10.

Jollimore, T. (2022) Impartiality. *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition), E.N. Zalta (ed.), https://plato.stanford.edu/archives/sum2022/entries/impartiality/.

Kamm, F.M. (2020) The use and abuse of the trolley problem: Self-driving cars, medical treatments, and the distribution of harm. In S. Matthew Liao (ed.), *Ethics of Artificial Intelligence*, 79–108. New York: Oxford University Press.

Kant, I. [1785]. *Groundwork of the Metaphysics of Morals*, in M.J. Gregor (ed.), *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy*, 38–108. Cambridge: Cambridge University Press (1996).

Kneer, M., Stuart, M.T. (2021) Playing the blame game with robots. In *Companion of the 2021 ACM/IEEE international conference on human-robot interaction*: 407–411.

Kochupillai, M., Lütge, C., Poszler, F. (2020) Programming away human rights and responsibilities? "The Moral Machine Experiment" and the need for a more "humane" AV future. *NanoEthics* 14(3): 285–299.

Lawlor, R. (2022) The ethics of automated vehicles: why self-driving cars should not swerve in dilemma cases. *Res Publica* 28(1): 193–216.

Lei, D., Chen, X., Zhao, J. (2018) Opening the black box of deep learning. *arXiv preprint arXiv:1805.08355.*

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., Legg, S. (2018) Scalable agent alignment via reward modeling: a research direction. *arXiv preprint. arXiv:1811.07871.*

MacAskill, W., Meissner, D., Chappell, R.Y. (2022) The rights objection. In R.Y. Chappell, D. Meissner, and W. MacAskill (eds.), *An Introduction to Utilitarianism*, https://www.utilitarianism.net/objections-to-utilitarianism/rights.

Montavon, G., Samek, W., Müller, K.R. (2018) Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73: 1–15.

Moor, J. (2009) Four kinds of ethical robots. *Philosophy Now* 72: 12–14.

Müller, V.C. (2014) Editorial: Risks of general artificial intelligence. *Journal of Experimental and Theoretical Artificial Intelligence* 26(3): 1–5.

Nascimento, A.M., Vismari, L.F., Queiroz, A.C.M., Cugnasca, P.S., Camargo, J.B., de Almeida, J.R. (2019) The moral machine: Is it moral? In *International Conference on Computer Safety, Reliability, and Security*, 405–410. Cham: Springer.

Nozick, R. (1974) *Anarchy, State, and Utopia*. New York: Basic Books.

Olsson, E. (2021) Coherentist Theories of Epistemic Justification. The Stanford Encyclopedia of Philosophy (Fall 2021 Edition), E.N. Zalta (ed.), https://plato.stanford.edu/archives/fall2021/entries/justep-coherence/.

Parfit, D. (2011) *On What Matters, Vol. 1*. Oxford: Oxford University Press.

Parfit, D. (1984) *Reasons and Persons*. Oxford: Clarendon Press.

Petersen, S. (2020) Machines learning values. In S.M. Liao (ed.), *Ethics of Artificial Intelligence*, 413–438. New York: Oxford University Press

Railton, P. (2020) Ethical Learning: Natural and Artificial. In S.M. Liao (ed.), *Ethics of Artificial Intelligence*, 45–78. New York: Oxford University Press.

Rawls, J. (1999) *A Theory of Justice, Revised Edition*. Cambridge: The Belknap Press of Harvard University Press.

Ross, W.D. (1930) *The Right and the Good*. Oxford: Clarendon Press.

Rudy-Hiller, F. (2018) The Epistemic Condition for Moral Responsibility. *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), E.N. Zalta (ed.) https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/.

Rummel, R.J. (2017) *Lethal Politics: Soviet Genocide and Mass Murder Since 1917*. New York: Routledge.

Russell, S. (2020) Artificial Intelligence: A Binary Approach. In S.M. Liao (ed.), *Ethics of Artificial Intelligence*, 327–341. New York: Oxford University Press.

Schwitzgebel, E., Garza, M. (2020) Designing AI with rights, consciousness, self-respect, and freedom. In S.M. Liao (ed.), *Ethics of Artificial Intelligence*, 459–479. New York: Oxford University Press.

Sherratt, Y. (2013) *Hitler's Philosophers*. New Haven: Yale University Press.

Shwartz-Ziv, R., Tishby, N. (2017) Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810.*

Simon, H.A. (1965) *The Shape of Automation for Men and Management*. New York: Harper & Row.

Sotala, K., Yampolskiy, R. V. (2014) Responses to catastrophic AGI risk: a survey. *Physica Scripta* 90(1), 018001.

Talbert, M. (2022) Moral Responsibility. *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), E.N. Zalta & U. Nodelman (eds.), https://plato.stanford.edu/archives/fall2022/entries/moral-responsibility/.

Wallach, W. Vallor, S. (2020) Moral Machines: From Value Alignment to Embodied Virtue. In S.M. Liao (ed.), *Ethics of Artificial Intelligence*, 383–412. New York: Oxford University Press.

Weng, Y.H., Chen, C.H., Sun, C.T. (2008) Safety intelligence and legal machine language: Do we need the three laws of robotics? In Y. Takahashi (ed.), *Service Robot Applications*, 195–214. Rijeka, Croatia: InTech.

Wiggers, K. (2022) DeepMind's new AI can perform over 600 tasks, from playing games to controlling robots. *TechCrunch*, https://techcrunch.com/2022/05/13/deepminds-new-ai-can-perform-over-600-tasks-from-playing-games-to-controlling-robots/.

Yampolskiy, R.V. (2020) On controllability of AI. *arXiv preprint. arXiv:2008.04071.*

Yarkoni, T. (2022) The generalizability crisis. *Behavioral and Brain Sciences*, 45.

Yudkowsky, E. (2011) Complex value systems in friendly AI. In J. Schmidhuber, K.R. Thórisson, and M. Looks (eds.), *Artificial General Intelligence: 4th International Conference*, AGI 2011 Proceedings: 388–393.

Zerilli, J., Danaher, J., Maclaurin, J., Gavaghan, C., Knott, A., Liddicoat, J., Noorman, M. (2021) *A Citizen's Guide to Artificial Intelligence*. Cambridge, MA: The MIT Press.