AGAINST MORAL DEFERENCE

WHY ARTIFICIAL MORAL AGENTS NEED NOT UNDERMINE PHRONESIS

JINGLONG YANG

§ 1: INTRODUCTION

In recent years, technological advances in artificial intelligence (AI) have accelerated the development of artificial agents. These agents—usually in the form of computer programs or robots—are designed to perform tasks in ways that emulate human action and reasoning. Their applications span a wide spectrum, including autonomous vehicles, humanoid companion robots, and algorithmic trading systems. As a result of rapid technological development, researchers in the field of machine ethics have begun to examine whether artificial agents might engage in moral deliberation or perform ethically significant actions. Drawing on concepts from human ethics and moral psychology, agents capable of ethical behavior are generally referred to as *Artificial Moral Agents* (AMAs).[1]

To be clear, not all ethical machines are AMAs. The taxonomy of ethical machines can be clarified by appealing to John Sullins's account of artificial agents, which distinguishes between *ethical impact agents* (EIAs), *artificial ethical agents* (AEAs), and *artificial moral agents* (AMAs).[2] EIAs are systems that merely generate outcomes with ethical consequences but lack any capacity for moral reasoning. For example, a self-driving car that unintentionally kills a pedestrian operates as an EIA. The next level of ethical agents is the AEAs. They are systems explicitly to incorporate ethical parameters—whether through rule-based architectures or machine learning algorithms. For example, an autonomous vehicle that chooses the least harmful crash option using

[1] Cervantes, et al. "Artificial Moral Agents," 501-532.
[2] Sullins, "Artificial Phronesis," 136-146.

utilitarian calculation would qualify as an AEA. However, AEAs remain morally passive, meaning they do not determine their own ethical standards but instead operate according to standards set by their human designers, who therefore retain ultimate moral responsibility for the agents' actions. Only at the level of AMA do we reach the threshold of systems capable of self-guided moral deliberation and context-sensitive ethical behavior akin to that of human agents.[3]

For the present purpose, this paper defines AMA as any *system that can choose its own ethical standards and act on its own ethical judgment in complex and novel situations without real-time human intervention.* The precise definition of an AMA remains contested, depending on what necessary conditions one takes to be constitutive of AMA creation. Proposed conditions include phenomenal consciousness (or its functional equivalent), rationality and moral competence, free will and autonomy, and moral responsibility.[4] Among these, consciousness seems to be the hardest to satisfy so far.[5]

> Nevertheless, even if consciousness proves to be a necessary criterion for AMA, its realization may not be impossible. According to Jeff Sebo and Robert Long's widely cited analysis of AI moral consideration, there exists a non-negligible probability (≥ 0.1%) that some AI systems will attain some degree of consciousness by 2030.[6]

Assuming such a precautionary stance, a pressing normative question arises: *should* humanity create AMAs if doing so becomes technically feasible?

Responding to this question, Aristotelian philosophers such as Nir Eisikovits and Dan Feldman have expressed concerns. Although their work does not address AMAs explicitly, their critique of advanced AI systems readily extends to them. Eisikovits and Feldman warn that such systems may erode human *phronesis*—our capacity for practical wisdom—by gradually displacing the exercise of moral judgment in everyday life.[7]

In this paper, I will contest Eisikovits and Feldman's pessimistic view and argue that, under appropriate conditions, AMAs can, in fact, facilitate cultivating rather than undermining human *phronesis*.

[3] Sullins, 139-140.
[4] Behdadi & Munthe, "A Normative Approach to Artificial Moral Agency," 195-218
[5] Himma, "Artificial agency, consciousness, and the criteria for moral agency," 19-29
[6] Sebo & Long, "Moral Consideration for AI Systems by 2030," 591-606
[7] Eisikovits & Feldman, "AI and Phronesis," 181-199

Eisikovits and Feldman's warning implicitly assumes that people will always, or at least, frequently enough, defer decision-making to highly intelligent AI systems, such as AMAs, to an extent that is damaging for our phronetic capacities. However, this assumption is far less secure than it appears. To demonstrate my views, I will first draw on Aristotle's account of responsibility in his *Nicomachean Ethics* to examine whether AMAs could satisfy the conditions for bearing responsibility. Eventually, failing to comply with the epistemic condition for responsibility renders AMAs incapable of being held accountable, despite their highly independent and autonomous capabilities. To prove why AMAs should never be trusted to make decisions independently, I will show in detail how they fail to fit within our current accountability mechanisms.

However, this limitation turns out to be beneficial. Precisely because AMAs cannot bear responsibility, humans cannot offload moral judgment onto them. Essentially, their unsuitability for bearing responsibility incentivises us to keep actively engaged in the deliberative process. Such active engagement is the ground for the human-AMA relationship to nourish and provide opportunities that not only preserve but also cultivate our phronetic abilities. In short, the creation of Artificial Moral Agents is not only permissible but desirable, given their potential to enhance human *phronesis*.

§ 2: CONCERNS ABOUT AMA

Despite the numerous threats posed by rapid AI development, many of them could be overcome. Some of the most prominent threats AI poses stem from methodological failures in the training and prediction stages.[8] Algorithmic bias is one of them. As Eisikovits and Feldman observe, AI's bias issues usually mirror and amplify the prejudices of its human designers. This kind of deficiency then contributes to entrenching the existing social and economic inequalities.[9] However, if these methodological defects can be remedied by advances in data curation, auditing, interpretability, and governance, a further question arises: "If algorithmic bias concerns about AI were eliminated, would there be anything left to worry about? To put it more sharply, if AI decisions became fairer than typical human decisions, would there be any residual discomfort with the technology?"

[8] Eisikovits & Feldman, 185
[9] Eisikovits & Feldman, 186

Other philosophers like Cheng-hung Tsai and Hsiu-lin Ku have advanced this question to a more radical level. They invite us to imagine a scenario in which not only algorithmic bias is resolved, but all technical aspects of AI, such as opacity, privacy, and hallucination, are refined to perfection,[10] just like the kind of artificial superintelligence (ASI) described by Nick Bostrom[11]—except in this case, the ASI does not pose existential risk. Then the question becomes: "If AI were to achieve technical perfection, would there be anything left to worry about with the technology?"

Eisikovits and Feldman point out that the worry lies in the deprivation of *phronesis*. Their normative concern can be illuminated through Aristotle's function argument in the *Nicomachean Ethics*:

1. The distinctive function of human beings is the excellent exercise of rationality, in which *phronesis* (practical wisdom) is a crucial part.

2. Sustained exercise of rationality cultivates virtues.

3. The possession and active exercise of virtues is constitutive of *eudaimonia* (happiness).

4. A being flourishes insofar as it actively and excellently performs its distinctive function.

5. Therefore, actively exercising rationality—such as *phronesis*—is a necessary condition for human flourishing and happiness.

If Aristotle is right, the exercise of rationality is not merely a means to flourishing but partly constitutive of it. This is the key of Eisikovits and Feldman's concern: by gradually displacing human decision-making—in domains such as hiring, resource allocation, and loan assessment—AI does not merely make certain tasks more efficient.[12] It deprives us of the very activity through which we become and remain good human beings. The threat, in other words, is not just to our competence but to our humanity.

AMAs may exacerbate such a normative concern. Continuing in the speculative sphere inherited from Tsai and Ku, a superintelligent AMA may be better than humans at making ethical decisions. From a utilitarian perspective, such superintelligent AMAs, capable of making fairer decisions faster and more consistently, are precisely

[10] Tsai & Ku, "Why AI may undermine phronesis," 3079-3086
[11] Bostrom, *Superintelligence: Paths, Dangers, Strategies*
[12] Eisikovits & Feldman, 189

where practical wisdom is most needed. Hence, people risk falling into moral deference with AMAs. If Eisikovits and Feldman are right about AI's undermining effect on human *phronesis*, does this mean we are doomed with the creation of AMAs? Should we prohibit such creations, even if they become technically feasible?

Not necessarily, but only if we scrutinize the assumption on which the threat rests. One key assumption is that people will defer moral decisions to AMAs. Although this assumption may be true to a certain extent, the scale of such moral deference is likely to be limited, given AMAs' inability to bear responsibility.

§ 3: ARISTOTELIAN CONDITIONS OF RESPONSIBILITY & AMA LIMITATIONS

§ 3.1: THE CONTROL CONDITION

Regarding the issue of responsibility, Aristotle offers us some useful insights in his *Nicomachean Ethics.* In Book III, he analyses responsibility in terms of two necessary conditions: the control condition and the epistemic condition. Essentially, an agent is responsible for his action if he has done it knowingly (the epistemic condition) and voluntarily (the control condition). Contemporary Aristotelian scholars often treat these two conditions as the basic criteria for attributing responsibility.[13]

The control condition states that an agent is responsible if the "principle of the action" is up to the agent to perform or refrain from it.[14] Philosophers often interpret this condition in terms of autonomy and freedom.[15] To be responsible, an agent must have adequate control over the deliberative process that leads to action; namely, the action must be voluntary, not forced.[16] When an action is forced, its cause lies outside of the agent's control, which may excuse the agent from taking responsibility. For example, if an AMA's decision-making process is externally overridden by a malicious hack, the "principle of action" would lie outside the system's control; hence, it might be

[13] Talbert, Moral Responsibility (Stanford Encyclopedia of Philosophy)
[14] Aristotle, NE III.1 1110a15–20
[15] Coeckelbergh, "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability," 2051-2068
[16] Aristotle, NE III.1 1110a1–5

excused from responsibility, just as a human acting under coercion might be. Essentially, the control condition places a heavy emphasis on individual control.

In the *Nicomachean Ethics*, Aristotle distinguishes autonomy from two levels: mere voluntariness and choice. He notes that children and even non-human animals act voluntarily insofar as their behavior originates from internal impulses.[17] But such behavior is not, strictly speaking, a choice. It is more of a reaction than an action. A choice, by contrast, involves deliberation about means within our power toward a certain end.[18] Choice requires a rational capacity to select among alternatives by considering how to best achieve some goals. This distinction opens conceptual space for AMAs. If advanced AMAs can evaluate alternative actions based on certain moral standards, they could be said to make choices rather than simply execute commands. The question is whether they could exercise rational control over means and practical ends in particular situations in a way that their "principle of actions" could be said to lie within themselves.

In theory, sufficiently sophisticated AMAs could satisfy this control condition, given enough technological advancements. Recall that an AMA is a *system capable of choosing its own ethical standards and acting on its own ethical judgment in complex and novel situations without real-time human intervention*. If such systems become technically feasible, they would, by definition, possess the deliberative capacities relevant to the control condition. Consider an AMA deployed to coordinate responses during an environmental crisis. Rather than executing a fixed decision rule, it needs to recognize multiple ethically salient considerations—intergenerational fairness, long-term ecological sustainability, and present human welfare—and deliberate about which should predominate. If new information emerges, the AMA should revise its practical priorities accordingly. This capacity to form, evaluate, and endorse practical ends in response to changing circumstances is precisely what distinguishes genuine deliberation from mere computation. In this sense, its "principle of action" lies within the system itself, thereby satisfying Aristotle's control condition.

Critics may object that AMAs cannot satisfy Aristotle's control condition because they cannot act contrary to their final ends (*telos*). At most, they can choose among means to ends that human designers have already fixed for them. In that sense, the AMA's principle

[17] Aristotle, NE III.2 1111b5–15
[18] Aristotle, NE III,2 1111b 5–1112a 15

of action remains external: its governing *telos* is not self-chosen, but imposed from outside.

This objection identifies a genuine limitation, but it mistakes the kind of control Aristotle regards as necessary for responsibility. Under his teleological metaphysics, Aristotle suggests that everything has a *telos* (or final end). Beings generally do not choose their final ends for themselves—including humans. We "wish" for our ends and "choose" among means.[19] Acorns do not choose if an oak tree is worth becoming, they just grow into it. Similarly, we do not deliberate about whether *eudaimonia* is worth pursuing but only how to pursue it. If responsibility required freely choosing one's final end, no human being could be responsible for anything. Aristotle avoids this absurdity by locating responsibility in our control over deliberation and action, not in originating our final ends. Hence, if the absence of freely chosen final ends disqualified an agent from responsibility, then humans and AMAs would fail together.

For this reason, the fact that AMAs do not choose their final ends does not by itself show that they fail Aristotle's control condition. In this respect, the difference between AMAs and human beings is less sharp than the objection suggests. Their final end would be fixed by their designers, just as our final ends may be fixed by our creator, whether that be God or Nature. Therefore, to say that humans fix the final ends of AMAs is only to say that humans may bear unavoidable responsibility for the creation and development of AMAs. It does not, however, follow that AMAs cannot have sufficient control over their own actions. What matters is whether an AMA has enough control over its practical ends and means for its principle of action to lie within itself in the relevant sense.

Although AMAs could not act contrary to their final ends, they may still act contrary to the subordinate practical ends assigned to them by humans. If they were truly intelligent enough, there is no obvious reason to believe that they cannot identify mistakes in a human operator's reasoning and propose alternative goals and corresponding means. Whether they should be given unrestricted authority to implement those plans is a separate question (more on this later). Still, that issue concerns the proper structure of human oversight, not whether the AMA exercises deliberative control. Consider again the environmental AMA. If it genuinely possessed ethical judgment of its own and operated by ethical standards it could apply autonomously, then a human instruction to reduce current carbon emissions might be

[19] Aristotle, NE III.2, 1112b11–17

treated not as a fixed directive but as a revisable practical end. The AMA could instead propose a different subordinate goal—say, investment in clean-energy research—if it judged that strategy better suited to the good of the environment. The key point is that, although AMAs would not choose their final ends, they could still possess sufficient control over their practical ends and means to satisfy Aristotle's control condition.

Another possible objection against AMAs satisfying the control condition is that they lack the ability to refuse tasks, and thus lack the autonomy that genuine control requires. Even if an AMA can propose different practical ends, once human operators insist on pursuing the end as they see fit, the AMA cannot refuse to comply.

It is certainly difficult to imagine why we would grant AMAs complete and unchecked authority over their actions. To do so would be highly risky, since it would amount to releasing an artificial agent without meaningful external constraint. For this reason, some limitation on refusal may be practically unavoidable. But for the same reason, the absence of a refusal capacity may reflect not a conceptual limitation in AMAs themselves, but a practical limit imposed by human designers. If we were ever willing to grant AMAs complete operational autonomy, then such systems could presumably acquire the ability to refuse tasks as well, and, in that respect, satisfy Aristotle's control condition more fully.

Although always being compelled to follow human instructions does weaken autonomy, it does not necessarily disqualify AMAs from satisfying the control condition. The fact that one can be forced does not prove one lacks autonomy altogether, but only that one lacks autonomy under certain situations. We certainly don't say that we, humans, are not voluntary agents simply because we can be coerced. Even without the full power of refusal, AMAs may still possess sufficient deliberative control over their practical ends and means to satisfy the control condition in the relevant sense. For example, they can communicate uncertainty when a situation exceeds their competence, flag potential conflicts between competing values, and propose alternative solutions. These are all meaningful exercises of practical deliberation. An AMA that proceeds blindly despite recognizing its limitations would indeed fail the control condition, but an AMA that does think critically and disagrees with its human operators is a different case. If after an AMA has told us its concern, explained why what we are asking it to do is bad or wrong, and proposed alternative solutions, it is still required to carry out the original instruction, then the resulting action may be

better understood as forced rather than fully voluntary. In this case, the AMA would be excused from responsibility, just as a human agent under coercion would be. After all, it would not choose that course of action if the decision were left to its own deliberation.

It is important to emphasize that this paper's focus is normative rather than technical. I am not claiming that current AI systems possess these capacities, nor am I speculating about when such capacities might be realized. The question under investigation is this: if technological development makes AMAs possible, should we create them? From the control condition alone, we can say that, at least, the technical advancements required for such deliberative capacities are not impossible. These advancements require nothing categorically different from incremental technological development. They do not demand, for instance, that a non-sentient system acquire consciousness. The control condition, as I have interpreted it, concerns only rational deliberation over means and practical ends—capacities that fall within the domain of sophisticated information processing rather than phenomenal experience.

However, the control condition alone does not qualify AMAs from bearing responsibility; they still face serious difficulties meeting Aristotle's epistemic condition.

§ 3.2: THE EPISTEMIC CONDITION

The epistemic condition is another foundational criterion that Aristotelian scholars recognize for attributing responsibility.[20] This condition requires the agent not be ignorant of what he is doing and of the relevant particulars of the situation.[21] More precisely, Aristotle distinguishes several important ways in which one can be ignorant:

> An agent acts involuntarily if he is ignorant of one of these particulars. They are: (1) who is doing it; (2) what he is doing; (3) about what or to what he is doing it; (4) sometimes also what he is doing with it—with the instrument, for example; (5) for what result—safety, for example; (6) in what way—gently or hard, for example."[22]

Ignorance of these particulars renders an action involuntary, thus not an appropriate target for responsibility. Rudy-Hiller translates this knowledge requirement to a bundle of awareness: to be responsible,

[20] Fischer & Ravizza, *Responsibility and Control*, 13
[21] Aristotle, NE III.1 1111a1–20
[22] Aristotle, NE III.1, 1111a3–5

an agent must be aware of the action itself, the instruments used, the consequences, and (according to some) the alternative options available.[23] Aristotle emphasizes that awareness of an action's consequences is the most crucial component.[24] Arguably, moral significance is an important consequence. Contemporary Aristotelian scholars usually call the capacity to discern what matters morally in a context-sensitive way *moral perception*.[25] [26]

To possess such perception, one must have an adequate understanding of virtue, which requires an appropriately formed emotional capacity. According to Aristotle, virtue is made of two components: virtue of knowledge and virtue of character.[27] Virtue of knowledge enables an agent to understand what ought to be done, while virtue of character ensures that the agent is affectively attuned to doing it in the right way.[28] On this view, correct moral perception not only requires correct judgment but also being educated to take pleasure and pain in the right things.[29]

These feelings do not merely accompany moral perception; they partly constitute it. A properly formed capacity for fear, shame, pity, pleasure, and pain allows an agent to register what is morally significant in a situation. We recognize threats as serious partly because we are capable of fearing what is genuinely fearful; we grasp another's suffering as morally salient partly because we are capable of being pained by it; and we become sensitive to moral failure partly because we are capable of shame. Without such affective states, an agent may still register the descriptive features of a situation, yet fail to grasp their ethical significance. In that sense, one who lacks the relevant feelings and emotions cannot fully perceive what moral perception requires.

Given this moral perception requirement, AMAs are unlikely to satisfy the epistemic condition. Unlike the control condition, which concerns capacities achievable through incremental technical progress, the epistemic condition involves something more fundamental. Until researchers solve the "hard problem of consciousness"[30] and discover

[23] Rudy-Hiller, The Epistemic Condition for Moral Responsibility (Stanford Encyclopedia of Philosophy)
[24] Aristotle, NE III.1, 1111a15–20
[25] McDowell, "Virtue and Reason," 331-350
[26] Liu, "Creating Character," 533-549
[27] Aristotle, NE II.1, 1103b15-19
[28] Aristotle, NE II.6, 1106b16-28
[29] Aristotle, NE II.3, 1104b3-1105a16
[30] Chalmers, "The Hard Problem of Consciousness"

how to instantiate phenomenal experience in artificial systems, AMAs will lack the qualitative mental states—fear, compassion, regret, empathy—necessary for genuine moral perception. An important consequence of this inadequacy is that AMAs do not experience pleasure and pain. But, according to Aristotle, repeated feedback of pleasure and pain grounds the habituation process that makes moral learning and character building possible.[31] Without lived experience of what feels good and what feels base, an agent cannot develop the emotional attunement that moral perception requires.

One might object that AMAs can simulate moral learning through reinforcement mechanisms, adjusting reward parameters in ways structurally analogous to how humans build character through habituation. However, this analogy is misleading. Human habituation is lived and felt; we become virtuous by reliably taking pleasure in virtuous actions and experiencing discomfort or shame when acting basely. Over time, these affective responses become second nature. A utility function might mimic the behavioral pattern of habituation, but it lacks the phenomenology that gives moral perception its depth. The difference is not merely quantitative but categorical. For example, consider a firefighter AMA deployed in disaster response. Such a system may assign high reward values to "saving human lives", thereby choosing actions that maximize survival. But the AMA has never lived. It doesn't understand what it feels like to have a family waiting at home, what it means to love someone, what is precious about being alive, or what is tragic about a preventable death. Its internal "awareness" is exhausted by data representations and optimization functions. This makes the AMA more akin to calculators that output "2" when given "1+1" than to human firefighters who risk their lives from a felt sense of duty and compassion. Such an AMA is not a moral agent exercising *phronesis* but simply a moral instrument extending the agency of its human operators.

Thus, AMAs are not fit to bear moral responsibility. They may perform morally correct outputs, but cannot genuinely understand why those outputs are morally significant. AMAs are, therefore, best understood not as responsible moral agents but as sophisticated moral instruments. Failing the epistemic condition disqualifies AMAs from being appropriate targets of praise or blame, thereby undermining any argument for entrusting them with independent moral decision-making. If AMAs cannot be responsible for their actions, then humans have strong reasons to retain moral authority rather than defer to these

[31] Aristotle, NE II.1, 1103a14-1103b25

systems. This observation will later prove to be significant for the question of whether AMAs threaten or facilitate human *phronesis*. But first, let us talk about why exactly humans have strong reasons to retain moral authority rather than defer to AMAs.

§ 4: PRACTICAL LIMITATION OF AMA DEVELOPMENT

§ 4.1: AMA LIMITATIONS IN ATTRIBUTION PRACTICES

Once the epistemic condition has failed, the basis for attributing responsibility is already severely weakened. This weakness becomes even clearer when we consider the practical aspects of our responsibility practices themselves. My key point is this: any attempt to attribute responsibility to AMAs will prove normatively thin, because it is difficult to see how AMAs could be appropriate bearers of responsibility in any sense meaningful to us.

In ordinary practice, responsibility attribution usually serves two aims: punishing the wrongdoer and providing redress for the victim. Accountability practices—punishment, blame, social condemnation—are morally significant because they impose a cost on the offender in proportion to their fault. Doing so acknowledges the victim's suffering and, to some extent, offsets that suffering. Punishments like imprisonment, fines, and public censure work precisely because they are experienced as painful or burdensome, damaging the offender's material circumstances, social standing, or sense of self. Similarly, redress is not merely a matter of compensation. Part of what victims seek is the recognition that the one who wronged them has been made answerable for that wrong. These practices presuppose that the responsible party can be meaningfully affected, both to himself and to those he has harmed.

That is precisely where AMAs fall short. Similarly to why they fail to pass the epistemic condition of responsibility, AMAs lack the kind of conscious moral understanding that would make praise or blame intelligible. Kenneth Himma has argued that a genuine moral agent must be an appropriate target of praise and blame.[32] This

[32] Himma, "Artificial agency, consciousness, and the criteria for moral agency," 19-29

requirement is not satisfied by the mere capacity to modify behavior in response to external inputs. It presupposes the ability to experience guilt, shame, suffering, or some analogous form of normative burden. Without phenomenal consciousness, AMAs do not experience pain as suffering, imprisonment as deprivation, social condemnation as a blow to their moral standing or self-conception. As a result, our ordinary accountability practices lose their normative force when directed at AMAs. What remains are mere technical interventions—shutting down the system, resetting its parameters, altering its utility functions. Such measures may be instrumentally useful for preventing future harm, but they do not amount to holding the AMA responsible in any morally substantial sense. Therefore, they are normatively thin.

Consequently, such technical interventions cannot provide meaningful redress for victims. When a human wrongs another, accountability practices communicate something morally important: that the wrong is recognized, that the wrongdoer is answerable, and that the victim's grievance has moral weight. None of these is adequately communicated to the victim by rebooting a server or adjusting a reward function. By analogy, "punishing" an AMA is comparable to kicking a rock that has rolled downhill and injured someone. The rock is insensible to the punishment; kicking it neither compensates the victim nor acknowledges the harm properly.

One might object that we already attribute responsibility to some non-conscious entities. Corporations, for instance, can be fined, sued, and publicly condemned. However, corporate responsibility is ultimately parasitic on human responsibility. When we hold a corporation accountable, we typically assume that the burdens of liability fall on human beings. The corporation serves as a juridical structure for assigning responsibility, but the normative force of accountability still traces back to persons capable of bearing its consequences. AMAs, by contrast, are not institutions composed of conscious members; they are technical systems with no conscious constituency to bear the burden of blame on their behalf.

This observation has led some ethicists to conclude that artificial agents should never be granted the status of responsible agents in the first place. Joanna Bryson, for instance, argues that robots "should be slaves"—that is, they should be treated as tools for which humans retain ultimate responsibility.[33] Her point is not that robots deserve slavery treatment, but that framing them as responsible

[33] Bryson, "Robots Should Be Slaves"

agents dissolves human accountability into technical artifacts that are structurally ill-suited to bear the moral weight we wish to place on them. As Andreas Matthias famously pointed out, if we pretend that advanced AI systems can be responsible, we risk creating responsibility gaps: situations in which serious harms occur but no one is genuinely answerable for them, because responsibility has been offloaded onto systems incapable of bearing it.[34]

Thus, AMAs may be important objects of regulation, but they are not fitting targets for moral responsibility. They can be managed when malfunctioning, but they cannot be blamed in any sense that satisfies the normative aims of our responsibility practices.

The fact that AMAs cannot bear responsibility has significant implications for how much trust we should place in them. If we were to allow AMAs to make ethical decisions independently, we would be entrusting beings incapable of accountability with matters of serious moral consequence. The bar for such trust, I believe, is extraordinarily high. It requires something approaching perfect reliability (say, an accuracy rate of about 99.99%).

The necessity for such stringent reliability becomes apparent when we recognize that complex ethical situations rarely involve a single decision. Typically, they require multiple layers of judgment: assessing the situation, identifying relevant considerations, weighing different alternatives, and adjusting as circumstances unfold. Even a highly reliable AMA will see its trustworthiness erode across successive decisions. Suppose an AMA achieves 95% accuracy on any single ethical judgment—an impressive figure by current standards. After five layers of decision-making, its cumulative reliability drops to approximately 77%. After ten layers, it falls below 60%. For an AMA to remain trustworthy across extended deliberative chains, it would need an accuracy rate approaching 99.99%—a threshold that may never be attainable in principle. Thus, even setting aside the question of responsibility, there are strong epistemic reasons for maintaining human authority over significant moral decisions.

[34] Matthias, "The Responsibility Gap," 175-183

§ 4.2: APPROPRIATE HUMAN-AMA COLLABORATION STRUCTURE

However, we should not disregard AMAs' highly capable capacities. If they can reliably perform morally correct actions, we should leverage this utilitarian advantage by having them as moral advisors, which can facilitate the cultivation of *phronesis* rather than undermine it.

Cheng-hung Tsai and Hsiu-lin Ku have already developed a promising response to Eisikovits and Feldman's concern that AI may undermine human *phronesis*. They propose a strategy called the *Principle of Epistemic Heed*, which holds that we should exercise our rational capacity as much as possible while paying careful attention to any superintelligent system's opinions. They emphasize that the crucial distinction is between heeding and deferring. To defer is to relinquish one's judgment in favor of another's; to heed is to attend carefully to another's opinion while retaining autonomous control over the decision-making process.[35] On this view, the proper relation between humans and highly intelligent systems is one in which such systems serve as advisors while we retain the final say. Thus, humans can benefit from an AMA's epistemic capabilities while still exercising their own practical judgment.

I wish to advance Tsai and Ku's position further. Their argument shows how epistemic heed can *preserve* opportunities for exercising phronesis; I contend that collaboration with AMAs can actually *cultivate* phronesis in ways that would not otherwise be available. There are two considerations to support this stronger claim.

The first is that highly capable AMAs can automate lower-order cognitive tasks that, while necessary, are often tedious and taxing. This automation frees us to focus on higher-order moral reasoning. For example, consider a hospital ethics committee evaluating organ allocation. An AMA could compile and synthesize medical data, flag inconsistencies in patient records, calculate survival probabilities, and identify applicable precedents from similar cases—tasks that are cognitively demanding but procedurally routine. By handling this informational groundwork, the AMA allows the committee to concentrate on harder questions: How should we weigh quality of life against length of life? What considerations of fairness apply when patients have unequal access to post-operative care? These are the kind

[35] Tsai & Ku, "Why AI May Undermine Phronesis," 3083

of questions that highly demand *phronesis*, and offloading lower-order tasks creates more cognitive space for engaging with them.

Second, having AMAs as moral advisors provides learning opportunities that enhance our judgment capacity. When a discrepancy arises between human and AMA judgment, we are compelled to engage closely with the AMA's reasoning, because, as I have argued, we remain the ultimate responsibility bearers. This kind of engagement typically requires higher-order thinking: challenging assumptions, weighing competing considerations, and arriving at a reflective judgment about which course of action is correct. Such deliberative work is difficult, but it is proportionally more rewarding for developing *phronesis*. If the AMA's recommendation proves to be correct, we gain an opportunity to reflect and refine our reasoning. Conversely, if we identify an error in AMA's reasoning, we have successfully engaged with a sophisticated deliberative process and exercised precisely the critical judgment that *phronesis* requires. In either case, so long as we heed rather than defer, we stand to benefit from the collaboration.

§ 4.3: REMAINING CONCERNS

Despite the advantages AMAs could bring, they also face serious drawbacks. Some of the major drawbacks can be best understood through Albert Borgmann's discussion of "focal practices". For Borgmann, focal practices are settled patterns of engagement with things that gather meaning and orient our lives—activities like running, gardening, or preparing a meal.[36] These practices cultivate human excellence precisely because they require sustained effort, skill, and attention; they unite means and ends, labor and enjoyment, in ways that develop our capacities. Borgmann warns that technological conveniences tend to threaten this unity by disburdening us of effort. Applied to AMAs, the concern is that automating aspects of ethical deliberation might similarly disburden us of the very engagement through which *phronesis* develops. If moral reasoning becomes something we receive rather than enact, we risk becoming passive consumers of ethical outputs rather than active practitioners of practical wisdom.

The first objection concerns passivity, illustrated by Borgmann's "easy chair problem." Borgmann observes that we often

[36] Borgmann, "Focal Things and Practices," 9-13

act against our better judgments because technology makes passivity easier.[37] For instance, when we come back home tired, we know a walk could refresh us, yet we may still choose to stay in a cozy chair with the television and a beer. Similarly with AMAs, we might know that genuine moral engagement is valuable, yet passively accept AMA recommendations out of convenience. Such passivity is indeed a problem, but may not be detrimental. Unlike the easy chair scenario—where the cost of passivity is low—moral decisions often carry significant stakes. Given that we remain the ultimate responsibility bearers, we have strong incentives to remain active in the decision-making process, thereby preserving the conditions under which *phronesis* can develop.

Another related objection concerns habituation. This objection holds that *phronesis* requires continuous habituation through repeated practice, not merely occasional engagement whenever discrepancies arise between human and AMA judgments. On this view, the "lower-order" tasks I propose to automate may themselves be essential to developing the stable dispositions that constitute practical wisdom. I acknowledge that there is genuine value in repeated practice of foundational tasks, especially their nurturing effect for higher-order reasoning. However, the concern is not unavoidable. Regarding children who are still developing their deliberative capacities, they should not be permitted to grow overly reliant on AMAs. Repeated practices have proven to be highly beneficial for children's learning in many aspects. Although the supervision required may be difficult to achieve, it is nevertheless addressable through education and regulation. As for adults with already-developed capacities, we can trust that they will remain fairly engaged with AMAs' reasoning, since, again, they bear ultimate responsibility for the resulting actions and decisions. Moreover, whenever adults feel that they need to sharpen their lower-order skills, they always have the option to do the work themselves and resume those focal practices.

The third objection concerns the opacity of AMA's reasoning. This is not an objection springing from Borgmann's discussion of focal practices; nevertheless, it is a strong challenge worth considering. The concern is this: if an AMA's deliberative process is not entirely transparent, can we engage with it meaningfully? I acknowledge that advanced AI systems may contain some opacity in their computational processes. However, they are not lacking in verbal deliberation ability. Current large language models are already capable of engaging in

[37] Borgmann, 14

conversation and explaining their reasoning when prompted. There is no reason to believe AMAs could not be paired with similar language abilities. As long as they can provide verbal explanations, there should be enough transparency for us to examine their thought processes. The goal is not to eliminate every opacity in the system, but simply to clarify aspects relevant to our decision-making context.

While concerns regarding passivity, habituation, and opacity are relevant, none of them proves fatal to my claim that AMA collaboration can cultivate human *phronesis*. The bottom line is this: as long as we remain the ultimate responsibility bearers, we are compelled to engage actively with AMAs rather than defer passively.

§ 5: CONCLUSION

The creation of Artificial Moral Agents is not destined to erode human *phronesis*. On the contrary, appropriately structured collaboration with AMAs can cultivate *phronesis*—provided that we understand the proper relationship between human and artificial moral agents.

That relationship is shaped by a fundamental asymmetry. AMAs, however sophisticated their deliberative capacities, do not satisfy Aristotle's epistemic condition for moral responsibility. Without phenomenal consciousness, they cannot experience the emotions that constitute moral perception. Hence, they should remain as moral instruments rather than moral agents. However, this apparent limitation turns out to be advantageous. Precisely because AMAs cannot bear responsibility, humans cannot offload moral judgment onto them. Any attempt to attribute responsibility to AMAs proves to be normatively thin. At the end of the day, we remain accountable for ethical decisions, and such accountability compels us to engage actively in the decision-making process.

This insight allowed me to advance beyond Tsai and Ku's Principle of Epistemic Heed. They showed that attending to AI suggestions while retaining autonomous judgment can preserve opportunities for exercising *phronesis*. I have argued for something stronger: that such collaboration can cultivate *phronesis* through two mechanisms—by freeing cognitive resources for higher-order reasoning, and by providing learning opportunities when human and AMA judgments diverge. In both cases, the collaboration demands

that we exercise precisely the capacities that practical wisdom requires: weighing competing alternatives, interrogating assumptions, and arriving at reflective judgments under conditions of genuine responsibility.

As for objections concerning passivity, habituation, and opacity, they are genuine but not fatal. After all, the stakes of moral decisions and our position as responsibility bearers incentivize our active engagement; developmental concerns can be addressed through education and regulation; and the opacity of AMA reasoning can be mitigated through their verbal deliberation capabilities.

I believe these findings have practical significance as AI systems increasingly enter domains requiring ethical judgment. The question is not whether to admit them into our moral lives but how to structure that collaboration wisely. This paper suggests that the answer lies in accountability: by ensuring that humans remain answerable for decisions made with AMA assistance, we create the conditions under which practical wisdom can flourish rather than atrophy. Essentially, the same feature that disqualifies AMAs from moral agency becomes the safeguard that preserves our own.

REFERENCES

Behdadi, Dorna, and Christian Munthe. 2020. "A Normative Approach to Artificial Moral Agency." *Minds and Machines* 30:195-218. https://doi.org/10.1007/s11023-020-09525-8.

Borgmann, Albert. 1984. "Focal Things and Practices." In *Technology and the Character of Contemporary Life: A Philosophical Inquiry*, 1-16. N.p.: University of Chicago Press.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. N.p.: Oxford University Press.

Bryson, Joanna. 2010. "Robots should be slaves." In *Natural Language Processing*. https://doi.org/10.1075/nlp.8.11bry.

Cervantes, Salvador, José A. Cervantes, Sonia López, Luis F. Rodríguez, Francisco Cervantes, and Félix Ramos. 2020. "Artifcial Moral Agents: A Survey of the Current Status." *Science and Engineering Ethics* 26:501-532. https://doi.org/10.1007/s11948-019-00151-x.

Chalmers, David. 2017. "The Hard Problem of Consciousness." In *The Blackwell Companion to Consciousness*, edited by Susan Schneider and Max Velmans. N.p.: John Wiley & Sons, Incorporated. https://doi.org/10.1002/9781119132363.ch3.

Coeckelbergh, Mark. 2020. "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability." *Science and Engineering Ethics* 26:2051-2068. https://doi.org/10.1007/s11948-019-00146-8.

Eisikovits, Nir, and Dan Feldman. 2022. "AI and Phronesis." *Moral Philosophy and Politics* 9 (2): 181-199. https://doi.org/10.1515/mopp-2021-0026.

Fine, Gail. 1996. *Aristotle: Introductory Readings*. Translated by Terence Irwin and Gail Fine. N.p.: Hackett Publishing Company, Incorporated.

Fischer, John M., and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. N.p.: Cambridge University Press.

Himma, Kenneth E. 2009. "Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?" *Ethics and Information Technology* 11:19-29. DOI 10.1007/s10676-008-9167-5.

Liu, Wei. 2012. "Creating Character: Aristotle on Habituation, the Cognitive Power of Emotion, and the Role of Prudence." *Frontiers of Philosophy in China* 7 (4): 533-549.

Matthias, Andreas. 2004. "The responsibility gap: Ascribing responsibility for the actions of learning automata." *Ethics and Information Technology* 6:175-183.

McDowell, John. 1979. "Virtue and Reason." *The Monist* 62, no. 3 (July): 331-350. https://doi.org/10.5840/monist197962319.

Rudy-Hiller, Fernando. 2018. "The Epistemic Condition for Moral Responsibility (Stanford Encyclopedia of Philosophy)." Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/moral-responsibility-epistemic/.

Sebo, Jeff, and Robert Long. 2023. "Moral Consideration for AI Systems by 2030." *AI and Ethics* 5:591-606. https://doi.org/10.1007/s43681-023-00379-1.

Sullins, John. 2021. "Artificial Phronesis: What It Is and What It Is Not." In *Science, Technology, and Virtues: Contemporary Perspectives*, edited by Emanuele Ratti and Thomas A. Stapleford, 136-146. N.p.: Oxford University Press.

Talbert, Matthew. 2019. "Moral Responsibility (Stanford Encyclopedia of Philosophy)." Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/moral-responsibility/.

Tsai, Cheng-hung, and Hsiu-lin Ku. 2025. "Why AI may undermine phronesis and what to do about it." *AI and Ethics* 5:3079-3086. https://doi.org/10.1007/s43681-024-00617-0.