AI and Ethics
https://doi.org/10.1007/s43681-024-00617-0

ORIGINAL RESEARCH

Why AI may undermine phronesis and what to do about it

Cheng-hung Tsai[1] · Hsiu-lin Ku[2]

Received: 27 May 2024 / Accepted: 10 November 2024

Abstract
Phronesis, or practical wisdom, is a capacity the possession of which enables one to make good practical judgments and thus fulfill the distinctive function of human beings. Nir Eisikovits and Dan Feldman convincingly argue that this capacity may be undermined by statistical machine-learning-based AI. The critic questions: why should we worry that AI undermines phronesis? Why can't we epistemically defer to AI, especially when it is superintelligent? Eisikovits and Feldman acknowledge such objection but do not consider it seriously. In this paper, we argue that there is a way to reconcile Eisikovits and Feldman with their critic by adopting the principle of epistemic heed, according to which we should exercise our rational capacity as much as possible while heeding a superintelligence's output whenever possible.

Keywords Artificial intelligence · Phronesis · Aristotelian · Superintelligence · Epistemic deference · AI reliance

1 Introduction

An agent with phronesis, or practical wisdom, is excellent at decision-making in human affairs.[1] Phronesis is not only considered as primarily valuable for human being, but also argued to be beneficial for the development of artificial intelligence (AI) [4, 10, 13, 16, 17, 22, 34]. However, a critical inquiry arises: does AI reciprocally contribute to the development of phronesis, or is there a possibility that AI not only fails to contribute but also presents a threat to phronesis? If AI may pose threats to phronesis, which is crucial for human flourishing, then such threats must be taken seriously and examined thoroughly to guide the responsible development and deployment of AI technologies.

In this paper we shall address why AI may undermine phronesis and what to do about it.[2] Specifically, our main focus will be on the latter part, because the former part has already been proposed by Nir Eisikovits and Dan Feldman in their "AI and Phronesis" [8], although we will elaborate further on their behalf.

Unlike the typical threats posed by AI to humans, which often stem from AI's own deficiencies, the threat underscored by Eisikovits and Feldman originates from human actions, with AI substantially exacerbating the issue. In Sect. 2, we outline Eisikovits and Feldman's worry regarding this threat

[1] According to Aristotle, a practically wise person is "able to deliberate nobly about what is good and beneficial for himself, not in particular respects, such as what conduces to health or strength, but about what conduces to living well as a whole" [1: 1140a]. Contemporary philosophers have offered more formalized accounts. For example, Stephen Grimm [11] argues that practical wisdom involves (1) knowing what is good or important for well-being, (2) knowing one's standing relative to what is good or important for well-being, and (3) knowing a strategy for obtaining what is good or important for well-being. Grimm's theory of wisdom is only *partially articulated*, while Cheng-hung Tsai [36–38] advances a *fully articulated* theory. Tsai argues that if a person S is wise, then (1) S knows that overall attitude success contributes to or constitutes well-being, (2) S knows what the best means to achieve well-being are, (3) S is reliably successful at acting and living well (in light of what S knows), and (4) S knows why he or she is successful at acting and living well. While a detailed definition is important, for ease of discussion in this paper, we may at times employ a simplified concept of wisdom, understood as knowing what matters, why it matters, and how to achieve it. Such knowledge, being practical in nature, is acquired by humans through practice and cannot be attained all at once.

[2] This expression imitates the title of the book by Mark Coeckelbergh [5].

✉ Cheng-hung Tsai
chtsai917@gate.sinica.edu.tw

Hsiu-lin Ku
hsiulinku@yahoo.com.tw

[1] Institute of European and American Studies, Academia Sinica, No. 128, Sec. 2, Academia Road, Nankang, Taipei 115, Taiwan

[2] Department of Philosophy, Chinese Culture University, No. 55, Hwa-Kang Road, Yang-Ming-Shan, Taipei 111, Taiwan

Published online: 16 December 2024

and formulate the principle underlying the worry, which we call the principle of epistemic fulfillment. In Sect. 3, we examine a way to dispel the worry through the principle of epistemic deference. In Sect. 4, we resolve the tension between the principles discussed in Sects. 2 and 3 by introducing the principle of epistemic heed. In Sect. 5, we suggest a procedure for putting the principle of epistemic heed into practice.

2 A worry: the principle of epistemic fulfillment

There are numerous threats posed by AI to human society, such as algorithmic bias and what Shannon Vallor [39] calls "acute technosocial opacity", which can be understood as the increasing opacity in our technological and social environment that makes it progressively harder to foresee the future amid rapidly changing technological, social, and contingent factors. One foreseeable prospect is that many of these threats can be overcome. However, once these threats, whether caused by deficiencies in AI technology or arising from the technosocial opacity, are resolved, does that mean there are no longer any threats posed by AI? This is the question that intrigued Eisikovits and Feldman: "If algorithmic bias concerns about AI were eliminated, would there be something 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: 187]?

Their question can be posed in a more radical manner. Algorithmic bias represents just one of the threats posed by AI to humans. The characteristics of these threats appear to stem either from flaws inherent in today's AI systems or from human factors. These issues could potentially be addressed in the future. Consequently, it seems that once AI achieves perfection, there would no longer be any threats to humans stemming from AI itself. By "perfection", we mean that the kind of artificial superintelligence (ASI) described by Nick Bostrom [3],[3] but with the distinction that this ASI would not pose the existential risks.[4] Thus, Eisikovits and Feldman's question can be framed as follows: If AI were to achieve perfection, would there be anything left to worry about with the technology?

Even if there were a *perfect* ASI, there would still be a threat posed by AI to humans. According to Eisikovits and Feldman, "if the person who has practical wisdom, the *phronimos*, is one who navigates particulars well, one who assigns appropriate weight to them based on context…, AI can emulate what that person does" [8: 191]. What is the problem if AI can emulate what a person with practical wisdom can do? Eisikovits and Feldman address their worry through Aristotle's function argument,[5] according to which human flourishing consists in the human function, and the distinctive human function is rational activity:

> Aristotle argues that human flourishing or *eudaimonia* is achieved through work– by practicing the capacities that, like our ability to make practical judgments, make us human. Now if AI, by replacing some of these practical judgments, results in us practicing less we will, through our engagement with this technology, become less of ourselves. What is at issue is the judge who… does less judging due to the introduction of sentencing guidelines, or the HR manager who does less hiring due to the algorithmic streamlining of her job. [8: 191]

To live well is to function well. To function well, for human beings, is to reason well. To reason well requires phronesis, which is a capacity to make excellent practical judgments. Against this background, Eisikovits and Feldman's worry can be stated as follows: "a key reason to worry about AI is that it undermines our capacity for practical judgment. By gradually taking over some of the contexts in which we exercise what Aristotle called phronesis, AI calls into question important aspects of our moral development" [8: 181].[6]

[3] Bostrom defines the concept of superintelligence as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" [3: 22]. He identifies at least three forms of superintelligence: "Speed superintelligence: A system that can do all that a human intellect can do, but much faster", "Collective superintelligence: A system composed of a large number of smaller intellects such that the system's overall performance across many very general domains vastly outstrips that of any current cognitive system", and "Quality superintelligence: A system that is at least as fast as a human mind and vastly qualitatively smarter" [3: 53–56].

[4] Bostrom thinks that "we cannot blithely assume that a superintelligence will necessarily share any of the final values stereotypically associated with wisdom and intellectual development in humans—scientific curiosity, benevolent concern for others, spiritual enlightenment and contemplation, renunciation of material acquisitiveness, a taste for refined culture or for the simple pleasures in life, humility and selflessness, and so forth" [3: 115–116]. That is, a superintelligence "could easily have non-anthropomorphic final goals, and would likely have instrumental reasons to pursue open-ended resource acquisition. If we now reflect that human beings consist of useful resources… and that we depend for our survival and flourishing on many more local resources, we can see that the outcome could easily be one in which humanity quickly becomes extinct" [3: 116]. Nevertheless, the focus of this paper is that, even if superintelligence itself is perfect—that is, both powerful and aligned with our values—there are still reasons for concern.

[5] See Eisikovits and Feldman [8: 187–188] for their explanation of the function argument. For a defense of Aristotle's function argument, see Whiting [40] and Korsgaard [19: ch.4].

[6] Eisikovits and Feldman's worry can be extended to various contemporary theories of wisdom that assert wisdom requires both learning and practice. In particular, it can be applied to the *skill model* of

Eisikovits and Feldman implicitly endorse a specific principle in expressing their worry over how AI undermines phronesis. By nature, "worry" stems from contemplating undesirable outcomes, which provoke feelings of distress and apprehension. If the undermining of phronesis by AI elicits such emotions, it likely indicates a misalignment with the underlying principle favored by Eisikovits and Feldman:

> **The Principle of Epistemic Fulfillment**
> The exercise and fulfillment of human's distinctive function of rationality is necessary for being human and achieving flourishing. We should therefore exercise and fulfill our rational capacity as much as possible.

This principle can be seen as a condensed form of the following argument: (P1) The exercise and fulfillment of human's distinctive function of rationality is necessary for achieving human flourishing. (P2) If something is necessary for achieving human flourishing, then we should pursue and prioritize that thing. (C) Therefore, we should exercise and fulfill our rational capacity as much as possible (from P1 and P2). What needs to be noticed is that this principle, on the surface, appears epistemic, but at its root, is ethical, for it treats human flourishing as basically or ultimately valuable.

Let us outline the factors causing AI to undermine phronesis and explain why this is a worry. Three factors work together. First, the development of rational capacity requires learning and practice to function well. Second, the decision-making processes of AI are increasingly mirroring our own, as demonstrated by the current advancements in AI, where deep learning algorithms are capable of accomplishing tasks traditionally carried out by humans. Third, there is a burgeoning reliance on AI for decision-making tasks among humans.[7] When these factors are considered collectively, it becomes apparent that, in decision-making contexts, humans are increasingly relying on AI (and eventually on perfect ASI) rather than on exercising and fulfilling their own rational capacity.[8] This trend, consequently, limits opportunities for individuals to develop their phronesis.[9] This outcome is particularly worrying due to the principle of epistemic fulfillment.

The worry raised by ASI in this paper is, in fact, a worry that AI—even without reaching the level of superintelligence—already poses to humans. At its core, any deviation from the principle of epistemic fulfillment constitutes this underlying worry. However, before reaching the stage of ASI, people often fail to fully grasp the far-reaching consequences of relying on AI for human well-being. Instead, they tend to attribute these problems to the immaturity of AI technology itself. In other words, when AI has not yet achieved the level of superintelligence, people may still recognize potential issues with the technology and choose to exercise their rational capacity to address them. As AI technology matures, however, ASI brings this worry into sharper focus, highlighting deeper concerns stemming from reliance on AI. This worry becomes increasingly apparent as the deviation from the principle of epistemic fulfillment becomes more pronounced. In essence, the emergence of ASI forces humanity to confront it more directly, particularly with regard to how AI might undermine phronesis and human well-being.

3 No worry: the principle of epistemic deference

In response to the worry that AI undermines phronesis, Eisikovits and Feldman suggest that "[a] technology that threatens to undo a foundational human capacity deserves closer moral scrutiny" [8: 197]. However, they do not elaborate much on what further actions should be taken. Yet, is it

wisdom, which argues that phronesis is a species of skill [2, 6, 32–34, 37]. According to Daniel Kahneman and Gary Klein, "a regular environment and an adequate opportunity to learn" are "preconditions for the development of skills" [15: 520]. So construed, phronesis, qua a skill, requires opportunity to learn. The rise of AI poses a threat to phronesis, as it diminishes our opportunity to learn.

[7] See Klingbeil et al. [18: 1], who note that "Where humans previously had to rely on their own expertise, nowadays complex decisions in many domains are supported by AI systems, such as hiring [20], investment advice [21], loan approval [28] and justice [24]".

[8] We are not suggesting that humans cannot rely on AI for decision-making, but rather that they should rely on it appropriately. How should this "appropriately" be understood? Most research on AI reliance defines *appropriate reliance* in decision-making as *adhering to correct AI advice* and *overriding wrong AI advice* (see [26]). However, current empirical studies have found that humans are often not able to achieve this type of appropriate reliance [9, 18] because humans tend to trust AI even when its advice is wrong. For an examination of the relationship between reliance and decision quality in AI-assisted decision-making, see Schoeffer et al. [27]. Nonetheless, our concern about appropriate reliance is broader: even if there is no wrong AI advice, complete reliance on AI still might be inappropriate.

[9] Previously, we mentioned that wisdom involves knowing what truly matters. Humans typically acquire this component of practical wisdom through lived experience and reflection, learning to discern what is most important in particular circumstances. Could AI potentially assist or even replace humans in acquiring this component of wisdom? Advanced AI systems can analyze enormous amounts of personal data—including social media interactions, official records, health metrics, behavioral patterns, personal preferences, browsing history, device data, shopping and transaction records, and public data sources—to detect subtle trends and provide outputs that might escape human notice. By comprehensively analyzing such life data, AI can prioritize individual values and deliver guidance not through personal reflection but via data-driven analysis. As AI continues to develop, it could potentially supplant some of the traditional human practices involved in acquiring practical wisdom.

conceivable that, in reality, no action is necessary, as there might not be any reason to worry in the first place?

The critic may argue that Eisikovits and Feldman's worry that AI undermines phronesis can be dispelled outright. But how is it possible to dispel their worry?

> Now a critic may plausibly object… along the following lines: Why is the concern about the atrophying of judgment such a big deal, why should we worry so much about our capacity for phronesis, especially if algorithmic replacements bring fairer, more equitable results for those who have been mistreated by our faulty, unfair, human, all too human judgments? Shouldn't we be concerned with better outcomes for those who were previously hurt by our misjudgments than about preserving the process of judging? Stated differently, perhaps the *virtuous* thing to do given the precarity and deficiency of our judgment is to gradually give it over to algorithms once it becomes clear they can do a better job than us? This is an important and powerful objection. [8: 192].

The critic argues that the worry is unfounded, asserting that it is ill-advised to rely on our own rational capacity in the era of AI or perfect ASI. In other words, the critic believes that, even if AI or perfect ASI could undermine phronesis, there should be no cause for worry.

We think that the critic implicitly endorses the following principle:

> **The Principle of Epistemic Deference**
> A future superintelligence occupies an epistemically superior vantage point: its beliefs are (probably, on most topics) more likely than ours to be true. We should therefore defer to the superintelligence's opinion whenever feasible. [3: 211][10]

Like the principle of epistemic fulfillment, the principle of epistemic deference[11] can be seen as a condensed form of an argument: (P1) A future superintelligence occupies an epistemically superior vantage point: its beliefs are more likely than ours to be true. (P2) If an entity's beliefs are more likely than ours to be true on most topics, then we should defer to this entity's opinion. (C) Therefore, we should defer to the superintelligence's opinion (from P1 and P2). Also like the principle of epistemic fulfillment, the principle of epistemic deterrence, on the surface, appears epistemic, but at its root, is ethical (or practical),[12] because it considers human flourishing to be ultimately valuable—it aims to minimize suffering and pain as a result of our erroneous judgments.

In response to the criticism, Eisikovits and Feldman make two claims. First, "if we accept the Aristotelian premise that the ability to make judgments is a large part of what people value about themselves (namely, *they don't exclusively value better outcomes*), then it is, at the very least, worth having a serious social discussion about whether the benefit of fairer outcomes is worth the cost of losing foundational capabilities" ([8: 192–193]; emphasis mine). Second, "if we accept the Aristotelian picture that tells us that the capacity for judgment is practice-dependent, that the ability to weigh particulars turns on having sufficient contexts in which to weigh them, the elimination of some of these contexts may diminish the capacity for judgment across the board" [8: 193].

These two responses, in our view, simply reiterate what we term the principle of epistemic fulfillment, which advocates for the active exercise and development of one's rational capacity as a moral imperative. These two responses appear to fall short of critically assessing, much less refuting, the rationale behind the principle of epistemic deference, which aims to prevent harm to an individual's flourishing caused by human errors, which is seen as a preferable outcome. Although the first response lightly touches on the rationale behind the principle of epistemic deference by suggesting that people "don't exclusively value better outcomes", it does not fully address or reject the rationale. Moreover, our daily lives are filled with examples of epistemic deference to domain-specific experts, such as doctors and scientists, who are often regarded as epistemic authorities.[13] This raises the question: does such deference undermine our phronesis? If it does not, then why resist accepting epistemic deference to a perfect ASI whose decision-making capacity surpasses ours?

[10] To align with the terminology used by Bostrom and other scholars, we employ terms such as superintelligence's "beliefs", "opinions", and "advice". However, we do not intend to attribute human-like traits to AI systems that do not possess them, nor do we believe that Bostrom intends to do so. To avoid the anthropomorphism fallacy noted by Hicks and Slater [12] and Placani [23], readers are encouraged to interpret any references to AI's "belief", "opinion", or "advice" as simply AI's "output" or to rephrase in a way that aligns with the context.

[11] "Epistemic deference" can be understood as "the phenomenon in which one person uses the opinions of another, either a real person or some idealized information source, as a model for what to believe" [14: 187].

[12] That is, to know what we want, or to identify what we value. Cf.: "…we may not know what we truly want, what is in our interest, or what is morally right or ideal. Instead of making a guess based on our own current understanding (which is probably deeply flawed), we would delegate some of the cognitive work required for value selection to the superintelligence. Since the superintelligence is better at cognitive work than we are, it may see past the errors and confusions that cloud our thinking" [3: 211].

[13] For a detailed account of the concept of epistemic authority, see Linda Zagzebski [41].

4 To worry or not to worry? The principle of epistemic heed

The discussion above reveals a fundamental tension between the principles of epistemic fulfillment and epistemic deference. Adherence to the former mandates the autonomous exercise of rationality, raising a worry in the age of AI, whereas the latter suggests relinquishing this exercise in favor of deferring to a superintelligence for the sake of better outcomes or utilities, thereby dispelling the worry. These principles, though seemingly divergent, both pivot around the axis of human flourishing and carry significant moral weight. The pivotal inquiry thus becomes: Between the principles of epistemic fulfillment and epistemic deference, which holds primacy? And under what conditions might one principle justifiably supersede the other?

The tension between these two principles can be seen as the tension between two ethical perspectives: Aristotelianism and utilitarianism. Aristotelianism focuses on the cultivation of virtues through habituation, while utilitarianism, as a form of consequentialism, emphasizes the outcomes of actions, aiming to produce the best possible consequences. Thus, the Aristotelian perspective aligns with the principle of epistemic fulfillment, highlighting the importance of exercising rational capacity and fulfilling human nature. In contrast, the utilitarian perspective aligns with the principle of epistemic deference, which favors deferring decision-making to agents—such as AI or superintelligence—that are perceived to yield optimal outcomes. From this perspective, the priority lies not in the act of rational engagement itself but in maximizing utility, minimizing error, and avoiding negative outcomes.[14]

So, it seems that our question of which principle–epistemic fulfillment or epistemic deference–holds primacy can be reframed as a debate between Aristotelianism and utilitarianism. However, investigating this debate does not alleviate the tension; instead, it sustains an ongoing philosophical discourse. This is particularly true in the realm of philosophy, where arguments can be endlessly constructed. Consequently, the tension between the principles of epistemic fulfillment and epistemic deference may remain unresolved for an extended period. This becomes problematic when *practical* decisions need to be made.

Having now established a clear understanding of the rationales behind the principles of epistemic fulfillment and epistemic deference, why not strive to satisfy both simultaneously? To achieve this, we propose the following principle:

> **The Principle of Epistemic Heed**
> We strive to fulfill our human nature by exercising our rational capacity, yet there is no guarantee against making erroneous judgments, which a perfect ASI, possessing an epistemically superior vantage point, can offer. We should therefore exercise our rational capacity as much as possible while heeding the perfect ASI's opinion whenever possible.

The principle of epistemic heed can be seen as a condensed form of an argument: (P1) We strive to fulfill our human nature by exercising our rational capacity, yet there is no guarantee against making erroneous judgments, which a perfect ASI can offer. (P2) If the perfect ASI can offer guarantee against making errors, then we should heed the ASI's opinion whenever possible. (C) Therefore, we should exercise our rational capacity as much as possible while heeding the perfect ASI's opinion whenever possible (from P1 and P2). As we can see, (P1) embraces the key notions encapsulated in the principles of epistemic fulfillment and epistemic deference, including the Aristotelian notion of human flourishing and the notion that a superintelligence possesses an epistemically superior vantage point.

The term "heed", as defined by dictionaries, signifies the act of paying careful attention to something, particularly advice or warnings. Importantly, paying attention does not equate to deference. This distinction allows for the independent exercise of one's rational capacity. Individuals can make decisions autonomously by considering the viewpoints of an ASI, rather than completely relinquishing control over their decision-making processes to the ASI.[15]

The concept of epistemic heed resolves the tension between the principles of epistemic fulfillment and epistemic deference by integrating the underlying rationales for both principles. The principle of epistemic fulfillment is rooted in the notion of satisfying our human nature through the exercise of our rational capacity. Meanwhile, the principle of epistemic deference focuses on maximizing utilities or outcomes, such as mitigating harms caused by human erroneous judgments, through epistemic deference to an ASI. The principle of epistemic heed illustrates how it is

[14] Ernest Sosa's concept of utilitarian questions provides insight into understanding utilitarianism as discussed here. According to Sosa, "Many questions call just for information. Take utilitarian questions, whether financial, legal, or medical. Answering such questions has a practical value fully realized with no need for deeper understanding" [30:4]. Additionally, he asserts, "utilitarian questions are properly answered with mere information acquired through sheer deference. But deeper choices require rational guidance beyond deference" [30: 6].

[15] Autonomous decision-making can be understood as a form of firsthand knowledge. See our discussion at the beginning of the next section regarding Sosa's distinction between firsthand and secondhand knowledge.

possible to exercise rational capacity while limiting human mistakes.

The advocate of the principle of epistemic deference might criticize by arguing that if an ASI's decision is indeed correct, why should we "heed" it rather than simply defer to it? The act of heeding the ASI's decision does not make the decision itself more correct; the quality of the decision remains the same. Therefore, the act of heeding, as embedded in the principle of epistemic heed, appears to be superfluous and a waste of cognitive resources, merely aimed at preserving the Aristotelian notion of human flourishing within the principle of epistemic fulfillment.

In our view, this criticism introduces a new criterion for deciding whether to accept the principle of epistemic heed or the principle of epistemic deference, namely, the conservation of cognitive resources. However, should the conservation of cognitive resources serve as a criterion to reject the principle of epistemic heed and support the principle of epistemic deference? Does the principle of epistemic deference meet the requirement of conserving cognitive resources? We think that in the context of achieving correct judgments and minimizing negative outcomes like suffering, the conservation of cognitive resources is not the primary consideration for the advocate of the principle of epistemic deference. If making correct judgments demands significant cognitive resources, the advocate of the principle of epistemic deference might argue that we should proceed regardless of the cognitive costs involved. Similarly, if achieving human flourishing demands significant cognitive resources, the advocate of the principle of epistemic heed might argue that we should proceed regardless of the cognitive costs involved. Thus, the inclusion of cognitive resource conservation as a criterion to reject the principle of epistemic heed may not be adequate. Consequently, the criticism can be nullified.[16]

5 Knowing how to epistemically heed

Our overall stance on the principle of epistemic deference isn't entirely negative; our stance aligns closely with Ernest Sosa's perspective, albeit Sosa isn't directly addressing the specific issue we're exploring in this paper (i.e., whether epistemically deferring to AI). If we define knowing through exercising our own rational capacity without deferring to others as *firsthand knowledge*, and knowing through deferring to others as *secondhand knowledge*, then, as Sosa suggests: "That is *not* to say that we cannot gain knowledge by just trusting someone. We can surely gain apt judgment that way. Secondhand knowledge is still knowledge, which we need not reject even in evaluative domains. The point is not to reject secondhand knowledge, but to embrace firsthand knowledge" ([29: 114]; see also [30: ch.1]).[17]

If we accept the principle of epistemic heed, a practical question might arise: How do we put the principle into practice? How do we exercise our rational capacity while also heeding the opinion of an ASI? To address this (although our subsequent discussion about such practices will remain abstract), let's first explore how the principles of epistemic fulfillment and epistemic deference might be put into practice. Those committed to the principle of epistemic fulfillment will exercise their rational capacity through methods such as deduction, induction, abduction, critical thinking skills, and more. Those committed to the principle of epistemic deference prioritize the ASI's opinions in a manner where, for instance, if the ASI has an opinion p, then this constitutes the reason for these individuals to believe p, replacing their other relevant reasons they have for believing p, rather than adding to those reasons. Moreover, if the ASI had belief q instead of p, these individuals would have reason to believe q instead of p.[18]

It is conceivable that those committed to the principle of epistemic heed will act similarly to those committed to the principle of epistemic fulfillment. It is also conceivable that those committed to the principle of epistemic heed will not act like those committed to the principle of epistemic deference. Instead, in scenarios where the opinions of an ASI are accessible, they heed, rather than merely defer to, its advice. The act of heeding includes comprehending the opinion offered by the ASI, understanding the reasoning behind it, and recognizing the data or logic that informed its conclusion.

[16] Here, we illustrate our point with an analogy. Humans possess various modes of transportation that enable them to move from point A to point B in the most efficient manner. If efficiency were always the primary consideration in moving from A to B, then people would not use their legs for this purpose. However, in certain contexts, individuals still opt to walk, jog, or run from A to B. This is because, in these situations, efficiency is not the primary consideration; gaining health benefits is, for example. We cannot criticize those who opt to walk, jog, or run from A to B based solely on efficiency.

[17] Why embrace firsthand knowledge? A succinct reply is: "Understanding through firsthand knowledge is salient for normative issues generally, and for moral issues more specifically. It is salient in the humanities, where we should and do often prioritize firsthand, nondeferential judgment" [30, p. 9].

[18] The way in which people committed to the principle of epistemic deference act, as conceived here, draws inspiration from and adapts Zagzebski's Preemption Thesis (i.e., "The fact that an authority has a belief p is a reason for me to believe p that replaces my other reasons relevant to believing p and is not simply added to them", [41: 107]) and the Content-independence Thesis ("an authoritative person or community's belief gives the subject a content-independent reason for belief. If the epistemic authority had believed a different proposition, the subject would have had reason to believe the other proposition instead", [41: 107]). Both theses, as articulated by Zagzebski, were influenced by the work of Joseph Raz.

Subsequently, practitioners of the principle of epistemic heed juxtapose the outcomes derived from exercising their rational capacity with opinions of the ASI. This involves evaluating both the parallels and variances between their judgments and the ASI's opinions. In situations where a discrepancy exists between their judgments and the ASI's opinions, the ASI's epistemically superior stance should be seriously considered.[19] Practitioners recognize that this superiority may arise from the ASI's capacity to process extensive datasets, its competence in calculating intricate probabilities, and its absence of cognitive biases to which humans are prone.[20] Therefore, practitioners will make decisions that take into account both their (initial) judgments and the ASI's opinions, with a tendency to heed the ASI's advice when its superiority is evident.

We view the process of epistemic heed outlined above as suggestive and not to be regarded as a fixed procedure. Indeed, we perceive the act of epistemically heeding ASI as akin to acquiring know-how,[21] representing a skill or multitrack disposition. In this regard, Aristotelians may find encouragement, as this scenario requires the full exercise of our practical rationality.

6 Conclusion

As noted above, Eisikovits and Feldman claim that "if we accept the Aristotelian premise that the ability to make judgments is a large part of what people value about themselves (namely, they don't exclusively value better outcomes), then it is… worth having a serious social discussion about whether the benefit of fairer outcomes is worth the cost of losing foundational capabilities" [8: 192–193]. The issue they bring up is crucial, and this paper goes further, not just focusing on fairer AI, but on AI that surpasses humans in all cognitive aspects (i.e., a perfect ASI). The aim is to address this issue by arguing that it is unnecessary to sacrifice one benefit for another; it is possible to achieve both. Once we grasp the principles and the rationales behind them, we can discern a path to fulfilling our human nature in the digital era.

Acknowledgements We sincerely thank the three anonymous referees for *AI and Ethics* for their valuable criticisms, comments and suggestions. This work is part of the project "The Expertise Theory of Wisdom Deepened, Extended, and Applied" funded by the National Science and Technology Council of Taiwan (NSTC 112-2410-H-001-030-MY3). The first author is also a member of the AI Governance Laboratory, IEAS, Academia Sinica.

Funding This research is supported by the National Science and Technology Council of Taiwan (Grant Number NSTC 112-2410-H-001-030-MY3).

Declarations

Conflict of interest The authors have no conflicts of interest to declare. We certify that the submission is original work and is not under review at any other publication.

References

1. Aristotle: The Nicomachean Ethics (trans. R. Crisp). Cambridge University Press, Cambridge (2000)
2. Baltes, P., Staudinger, U.: Wisdom: a metaheuristic (pragmatic) to orchestrate mind and virtue toward excellence. Am. Psychol. **55**(1), 122–136 (2000)
3. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford (2014)
4. Casacuberta, D.: The quest for artificial wisdom. AI Soc. **28**, 199–207 (2013)

[19] In situations where there is a discrepancy between human judgments and the ASI's opinions, it suggests that human judgment may have erred. Such discrepancies or errors serve as valuable learning experiences, providing opportunities to reflect on the reasons behind the initial judgment and identify the factors that led to the error, thereby enhancing the capacity for sound judgment. Thus, the principle of epistemic heed neither disregards the importance of making mistakes nor limits the opportunities for humans to practice and develop their judgment through these mistakes. We appreciate the reviewer for raising this question.

[20] According to Bostrom, the superiority of ASI does not arise out of thin air or from science fiction; rather, it stems from conditions that are superior to those of humans. In a section titled "Sources of advantage for digital intelligence" in his *Superintelligence* [3: 59–61], Bostrom states, "the hardware advantages [of digital minds] are easiest to appreciate" (p.59). These advantages include capabilities that surpass human limitations, such as the speed of computational elements, internal communication speed, number of computational elements, storage capacity, reliability, lifespan, sensors, etc. Additionally, he mentions that "Digital minds will also benefit from major advantages in software" (p.60), in areas such as editability, duplicability, goal coordination, memory sharing, and new modules, modalities, and algorithms.

[21] Regarding contemporary philosophical discussions on know-how, refer to the classic discussion by Ryle [25], as well as intellectualism about know-how such as Stanley and Williamson [31], anti-intellectualism about know-how like Dreyfus [7], and hybrid accounts of know-how such as Tsai [35], among others.

5. Coeckelbergh, M.: Why AI Undermines Democracy and What to Do About It. Polity Press, Cambridge (2024)
6. De Caro, M., Vaccarezza, M., Niccoli, A.: Phronesis as ethical expertise: Naturalism of second nature and the unity of virtue. J. Value Inq. **52**, 287–305 (2018)
7. Dreyfus, H.: Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action. Oxford University Press, Oxford (2014)
8. Eisikovits, N., Feldman, D.: AI and phronesis. Moral Philos. Politics **9**(2), 181–199 (2022)
9. Fok, R., Weld, D.S.: In search of verifiability: explanations rarely enable complementary performance in AI-advised decision making. AI Mag., online, 1–16 (2023).
10. Goertzel, B.: Artificial Wisdom. IEET, Institute for Ethics and Emerging Technologies (2008).
11. Grimm, S.: Wisdom. Australas. J. Philos. **93**(1), 1–16 (2015)
12. Hicks, M., Humphries, J., Slater, J.: ChatGPT is bullshit. Ethics Inf. Technol. **26**, 38 (2024)
13. Jeste, D., Graham, S., Nguyen, T., Depp, C., Lee, E., Kim, H.-C.: Beyond artificial intelligence: Exploring artificial wisdom. Int. Psychogeriatr. **32**(8), 993–1001 (2020)
14. Joyce, J.: Epistemic deference: the case of chance. Proc. Aristot. Soc. **107**, 187–206 (2007)
15. Kahneman, D., Klein, G.: Conditions for intuitive expertise: a failure to disagree. Am. Psychol. **64**(6), 515–526 (2009)
16. Karlan, B., Allen, C.: Engineered wisdom for learning machines. J. Exp. Theor. Artif. Intell. **36**(2), 257–272 (2024)
17. Kim, T., Mejia, S.: From artificial intelligence to artificial wisdom: what socrates teaches us. Computer **52**, 70–74 (2019)
18. Klingbeil, A., Grützner, C., Schreck, P.: Trust and reliance on AI—an experimental study on the extent and costs of overreliance on AI. Comput. Hum. Behav.. Hum. Behav. **160**, 108352 (2024)
19. Korsgaard, C.: The Constitution of Agency: Essays on Practical Reason and Moral Psychology. Oxford University Press, Oxford (2008)
20. Li, L., Lassiter, T., Oh, J., Lee, M.K.: Algorithmic Hiring in Practice: Recruiter and HR Professional's Perspectives on AI Use in Hiring. Proceedings of the 2021 AAAI/ACM conference on AI. Ethics, and Society (2021).
21. Lourenço, C.J., Dellaert, B.G., Donkers, B.: Whose algorithm says so: the relationships between type of firm, perceptions of trust and expertise, and the acceptance of financial robo-advice. J. Interact. Mark. **49**, 107–124 (2020)
22. Marsh, S., Dibben, M., Dwyer, N.: The Wisdom of Being Wise: A Brief Introduction to Computational Wisdom. In Habib, S., Vassileva, J., Mauw, S., Muhlhauser, M. (eds.) Trust Management X. IFIPTM 2016. IFIP advances in information and communication technology, vol. 473, pp. 137–145 (2016).
23. Placani, A.: Anthropomorphism in AI: hype and fallacy. AI Ethics (2024). https://doi.org/10.1007/s43681-024-00419-4
24. Re, R.M., Solow-Niederman, A.: Developing artificially intelligent justice. Stanford Technol. Law Rev. **22**, 242 (2019)
25. Ryle, G.: The Concept of Mind. University of Chicago Press, London (1949)
26. Schemmer, M., Kuehl, N., Benz, C., Bartos, A., Satzger, G.: Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. Proceedings of the 28th International Conference on Intelligent User Interfaces, pp. 410–422 (2023).
27. Schoeffer, J., Jakubik, J., Vössing, M., Kühl, N., Satzger, G.: AI Reliance and Decision Quality: Fundamentals, Interdependence, and the Effects of Interventions" (No. cekm9). Center for Open Science (2024).
28. Sheikh, M.A., Goel, A.K., Kumar, T.: An approach for prediction of loan approval using machine learning algorithm. In: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (2020).
29. Sosa, E.: Firsthand knowledge and understanding. In: Grimm, S. (ed.) Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology, pp. 109–122. Oxford University Press, Oxford (2019)
30. Sosa, E.: Epistemic Explanations. Oxford University Press, Oxford (2021)
31. Stanley, J., Williamson, T.: Knowing how. J. Philos. **98**(8), 411–444 (2001)
32. Swartwood, J.: Wisdom as an expert skill. Ethical Theory Moral PractPract. **16**, 511–528 (2013)
33. Swartwood, J., Tiberius, V.: Philosophical foundations of wisdom. In: Stenberg, R., Gluck, J. (eds.) The Cambridge Handbook of Wisdom, pp. 10–39. Cambridge University Press, Cambridge (2019)
34. Tsai, C.: Artificial wisdom: a philosophical framework. AI Soc. **35**(4), 937–944 (2020)
35. Tsai, C.: Beyond Intuitive Know-How. Phenomenology and the Cognitive Sciences (2022a).
36. Tsai, C.: Practical wisdom, well-being, and success. Philos. Phenomenol. Res. **104**(3), 606–622 (2022)
37. Tsai, C.: Wisdom: A Skill Theory. Cambridge University Press, Cambridge (2023)
38. Tsai, C.: Phronesis and emotion: the skill model of wisdom developed. Topoi Int. Rev. Philos. **43**, 1011–1019 (2024)
39. Vallor, S.: Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press, Oxford (2016)
40. Whiting, J.: Aristotle's Function Argument: A Defense. Ancient Philosophy 8(1): 33–48. Reprinted in J. Whiting, Living Together: Essays on Aristotle's Ethics. Oxford University Press, 2023, Ch.1 with a Postscript (1988).
41. Zagzebski, L.: Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief. Oxford University Press, Oxford (2012)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.