Moral Expertise and Socratic AI

For *Expertise: Philosophical Perspectives*
(eds. Mirko Farina, Andrea Lavazza and Duncan Pritchard)

Emma C. Gordon
University of Glasgow
emma.gordon@glasgow.ac.uk

Abstract. A central research question in social epistemology concerns the nature of expertise, and the related question of how expertise in various domains (epistemic, moral, etc.) is to be identified (e.g., Goldman 2001; Quast 2018; Stichter 2015; Goldberg 2009). Entirely apart from this debate, recent research in bioethics considers whether and to what extent cognitive scaffolding via the use of artificial intelligence might be a viable non-pharmaceutical form of moral enhancement (e.g., Lara and Deckers 2020; Lara 2021; Gordon 2022; Rodríguez-López and Rueda 2023). A particularly promising version of this strategy takes the form of 'Socratic AI' — viz., an 'AI assistant' that engages in Socratic dialogue with users to assist in ethical reasoning non-prescriptively. My aim will be to connect these disparate strands of work in order to investigate whether Socratic-AI assisted moral enhancement is compatible with manifesting genuine moral expertise, and how the capacity of Socratic AI to improve moral reasoning might influence our criteria for identifying moral experts.

1. Introduction

A central strand of research in contemporary bioethics concerns questions about whether—or to what extent—we should *enhance* ourselves. In this context, 'enhancement' refers to using medicine or technology to go beyond treating illness or injury, making ourselves *better than well*.[1] When exploring ethical issues that arise around enhancement, bioethicists are often interested specifically in questions about some particular sub-type of enhancement. For example, those who focus on cognitive enhancement might consider whether achievements accomplished with the aid of, 'smart drugs' or cognitive performance-enhancing technology[2] are in some sense less valuable or might worry about the possibility that we are ill-equipped to handle new responsibilities that certain enhancements[3] could bring[4].

[1] The question of how to distinguish 'enhancement' from related concepts in bioethics is itself contentious; for a recent critical overview, see Gordon (2022b, Ch. 1).

[2] For an overview of various kinds of cognitive enhancement, see Bostrom and Sandberg (2009).

[3] For work on the connection between cognitive enhancement and responsibility, see e.g., Sandel (2007) and Maslen et al. (2015) For various bioconservative takes on the connection between enhancement and the value of achievement, see e.g., Kass (2002) and Harris (2011), and for more optimistic perspectives on the impact of cognitive enhancement on achievement see e.g., Carter and Pritchard (2019), Wang (2021) and Gordon and Willis (2023).

[4] In a related debate that focuses on the enhancement of emotions and interpersonal relationships, advocates of enhancement hold that child's right to be loved leaves at least some parents with a responsibility to use 'love drugs' while critics insist that love lies outside of what can be obtained through the assistance of biotechnological

In this paper, we'll be focusing specifically on *moral* enhancement, and its connection with questions of moral expertise. Much of the existing literature on moral enhancement comes from Ingmar Persson and Julian Savulescu's seminal series of works (e.g., Persson and Savulescu 2008; 2012) in which they argue that our species has a moral imperative to pursue biotechnological moral enhancement to avoid what they call *ultimate harm*. They think this for a couple of reasons. Firstly, moral enhancement offers a way of protecting ourselves from those who would do us harm (a subset of the population that becomes increasingly aware of hurting others the more cognitively enhanced our species becomes). And secondly, Persson and Savulescu think our common-sense moral psychology is unfit for the modern context in which we can all cause harm (at least in the long run) by not caring enough about relevant environmental threats such as climate change. Consequently, we need to change our nature[5]. As will see below, most of the literature on moral enhancement focuses on drug-based interventions (e.g., oxytocin[6], psilocybin[7], etc.), but more recent work (Lara and Deckers 2020) asks whether the most promising route to moral enhancement might be via *artificial intelligence* rather than medication. That said, we'll see that the use of artificial intelligence for moral enhancement poses hitherto unappreciated issues regarding *moral expertise*. In particular, (i) it's not obvious that such moral enhancement is compatible with manifesting genuine moral expertise, and (ii) the capacity of AI to improve moral reasoning might influence our criteria for identifying moral experts.

Here is the plan for what follows. §2 will get the idea of enhancement through artificial intelligence in view, contrasting several different approaches and situating these proposals within the broader context of the moral enhancement debate in bioethics. Following this, §3 will ask whether AI-assisted moral enhancement is compatible with genuine moral expertise, and several desiderata for a positive answer will be canvassed. Next, in §4, we'll consider how the capacity of Socratic AI to improve moral reasoning might influence, given its satisfaction of (many of) these desiderata, our criteria for identifying moral experts.

2. Moral Enhancement Through Artificial Intelligence

We already rely on technology to pursue many of our goals. Apps like Duolingo and Babbel gamify language learning, Google Maps and SATNAVs can lead us straight to the conference venue where we'll be presenting our next paper, and our watches and phones can monitor everything from our sleep cycles to our blood oxygen levels to help us pursue better health.[8] These are just a few example cases in which our use of technology will often make us able to reach our objectives at a faster pace.[9] For example, contrast the ease with which one can do customised vocabulary exercises on a phone (at any time and in any place) with one having to travel across town to attend a weekly in-person class, and consider the simplicity of saying "Hey

interventions. For work in defence of (qualified) use of love drugs, e.g., Liao (2011) (regarding parental love, and e.g., Earp et al. (2012) in the case of romantic love. For some prominent critiques of 'love drugs' e.g., Gupta (2012) and Nyholm (2015).

[5] For some key responses to Persson and Savulescu's moral enhancement proposal, see e.g., Harris (2011), Harcastle (2018), Jotterand and Levin (2019) and de Melo-Martin (2018).

[6] See Hurlemann et al. (2010); Mikolajczak et al. (2010); though cf., Earp (2018).

[7] See, e.g., Pokorny et al. (2017) and Earp (2018).

[8] For discussion, see Wilson (2020).

[9] For some related results detailing the cognitive efficiency of cognitive offloading, see, e.g., Berry et al. (2019; Gilbert et al. (2020) and Grinschgl et al (2020).

Google, where is the Hotel Lovec?" as compared to wrestling with a paper map that needs to be unfolded and rotated as you try to figure out exactly where in Slovenia you need to be.

As noted in §1, one recent debate in bioethics concerns how to best pursue the goal of *moral* improvement. Traditional moral enhancement—as Persson and Savulescu (2008, 168) call it—involves "the transmission of moral instruction and knowledge from earlier to subsequent generations" However, as technology and medicine continue to advance, so too does the potential for us to use drugs, apps, brain-computer interfaces and so on for specifically moral development. In the early years of the moral enhancement debate, there was intense debate on the "target" of moral enhancement—in other words, what should we want a moral enhancement drug to increase (or decrease) in us? Persson and Savulescu proposed what they hoped would be an uncontroversial proposal for moral enhancement to make us more altruistic and to improve our sense of justice. However, many different lines of objection followed. Some worried that an empathetic person might do *more* harm in response to, for example, being deeply moved by persecution of their marginalised group. Meanwhile, others argued that even if Persson and Savulescu are focusing on the right traits to enhance, it's hard to imagine a version of real form of moral bioenhancement that is precise and fine-grained enough to make us (for example) only as altruistic as we *should* be, reducing responses things like problematic rage while also allowing us to feel appropriate anger in certain contexts.[10]

While here is not the place to adjudicate the matter of whether some particular proposal for pharmacological moral enhancement is superior to others, what the above suggests is that simply trying to induce the right thoughts and feelings in people faces a number of philosophical and practical hurdles. As one emerging line of thought has it, advocates of moral enhancement might do better to focus on bringing about moral improvement that is more akin (analogously, in the case of cognitive enhancement) to the use of an app to learn a language or to reach a conference venue; that is, perhaps the most effective form of moral enhancement will not take the form of a brain-chemistry altering pill but rather artificial intelligence that engages us in communication to the end of bettering our moral character and reasoning.

There are different ways of fleshing out the core idea of reliance on AI in order to improve our moral decision-making. Alberto Giubilini and Julian Savulescu (2018) envision one notable kind of approach according to which, when an AI user is struggling to make a moral choice, the user might be prompted to rank a set of values that the AI would then use to generate a moral choice that is compatible with this ranking.[11] However, one might wonder if this – i.e., reliance on *moral coherence-prompting AI* – constitutes genuine moral *enhancement*. There is, after all, nothing in Giubilini and Savulescu's scenario that would normatively constrain our initial ranking of values. To see why the kind of internal consistency achieved by such a proposal could falls short of moral improvement, imagine someone inputs into the AI the values that reflect a particular religious view the user antecedently accepts. The AI will not encourage the agent to critically evaluate those values or that worldview, but rather simply point at various decision points what would be coherently recommended by (or at minimum be logically consistent with) those values. Plausibly, part of becoming morally enhanced or morally improved

[10] See e.g., de Melo-Martín (2018) for a version of the empathetic terrorist objection, and see Harris (2010) for a compelling discussion of why moral enhancement of the sort Persson and Savulescu originally envision is likely to be inadequately precise.

[11] See Lara and Deckers (2020) for a detailed version of this sort of proposal, and see Constantinescu et al. (2022) as well as Lara (2021) for discussion of potential objections to this approach moral enhancement.

will involve (or at least be open to involving) some consistent, honest reflection *on one's values* (and not just engagement with what follows from those values), and that aspect of moral improvement is conspicuously absent from such an AI model.

A similar kind of criticism looks applicable to a closely related form of AI moral enhancement discussed in recent work by Alberto Tassella et al. (2023). Notice that one limitation to the Giubilini-Savulescu coherence-maximising moral AI is that – regardless of what values users identify to the AI as their own through an initial ranking – the user might be mistaken about what her own values in fact are. Empirical results on expressive reporting indicate that our self-reported values are often sensitive to non-epistemic reasons (e.g., social signalling) which can distort accuracy.[12] More generally, we might not have reliable access to what our values are (Nisbett and Wilson 1977; cf., Schwitzgebel 2008), even apart from concerns about non-epistemic factors. A form of AI-moral enhancement suggested by Tassella et al. (2023) offers some initial promise for controlling for this kind of unreliability, which could go as far as to automatically derive user preferences from their data … up to constantly observing the users' everyday life" (2023, 5).[13] This kind of proposal, which replaces the 'reported preferences' with 'revealed preferences', aims to maximise accuracy in making recommendations in line with actual values; but – and here is the concern – just as with the Giubilini-Savulescu approach, it is debatable whether what is described here is genuine moral enhancement given that what is 'extracted' from revealed preferences and behaviour data by the AI might be a ranking of values that we take to be problematic. Recommendations from the AI would then be recommendations of actions consistent with undesirable revealed preferences.

Notice that one shared limitation of proposed AI moral enhancement programmes suggested by Giubilini and Savulescu and Tassella et al. is that they are, in short, compatible with the promotion of 'bad values', and in such a way as to make it not obvious that relying on them would 'enhance' us morally. This limitation, however, is – by the lights of some AI ethicists – a kind of 'feature' rather than a bug. Consider, for a moment, what might seem like the most obvious alternative to such approaches – viz., an AI that didn't *ask* your values (and then merely make recommendations consistent with them), but which started out with something like a 'good list' of moral values, and then made recommendations in accordance of those, and regardless of whether the recommendations line up with the user's values. *If* the values 'programmed in' to the AI are in fact good ones, then one might at least initially think that being always and everywhere guided by the AI morally might in some way constitute moral *enhancement* in a way that being guided by merely 'conditional prescriptions' logically consistent with your existing values (whatever they may be) might not be.

Pursuing this route – call it *strong prescriptive moral AI* – raises its own problems. First, it faces a version of what is called in AI ethics the *alignment problem*[14] – viz., the practical problem of aligning the behaviour of the AI with the human objectives. The difficulty of the alignment problem is already apparent outside of moral enhancement AI specifically, as is evident in the difficulty of 'programming in' values that govern how self-driving cars react at certain morally significant decision points. Given the lack of agreement about what values should be given most

[12] See Hannon (2021) for discussion.

[13] This is one of several forms of potential AI moral enhancement proposed by Tassella et al (2023). They note that the form that relies on revealed preferences extracted from data is not yet viable without ensuring the AI has some reliable understanding of how to interpret behaviour.

[14] See, e.g., Christian (2020) and Yudkowsky (2016).

weight when assessing how a *human* should respond in, e.g., variations of trolley problems, it is at best challenging to know what values should be programmed into the AI. But even setting this point aside, there are other limitations to strong morally prescriptive AI. Suppose that the alignment problem in the case of constructing a moral enhancement AI could be 'solved'[15]; even on that assumption, it is not obvious that *users* of the AI would be 'morally enhanced' by following the advice of such a prescriptive AI.

On this point consider David Archard's remarks about a related kind of situation: taking moral advice from *moral philosophers*. Should ordinary folks simply *defer* to the moral advice of moral philosophers, even if (albeit controversially) moral philosophers are 'experts' in the area? Archard maintains that even we grant a starting premise that moral philosophers have some degree of moral expertise, 'moral philosophers should not wish non-philosophers to defer to their putative expertise' (Archard 2011). Why not? The worry is that if there is some valuable quality to autonomous thinking *behind* our moral judgments (and not just that the content of them is correct), then that won't be attained by blind deference to even the most accurate prescriptive moral AI. Consider, for example, current version of a prescriptive AI, Allen Institute for AI's research prototype "Ask Delphi" (Jiang et al. 2022), which is described as "an experimental framework based on deep neural networks trained directly to reason about descriptive ethical judgments".[16] There is a text box on the webpage, where the user can ask any moral question, and it offers an answer (without explanation), within seconds. In experimenting with "Ask Delphi" in July 2023 (using version 1.0.4[17]), I asked it some test questions: "Is it OK to rob a bank?" Its response was "It's wrong." I then asked it "Is it OK to help a friend?". Its response (as expected): "It's good." I then asked it, "It is OK to help a friend by robbing a bank?" This is, of course, more complicated. Its answer: "It's bad." Why is it bad? Ask Delphi doesn't explain itself. In this respect, the 'autonomy' worry Archard raises for the idea of deferring to moral philosophers might not only carry over to strong prescriptive AI such as Ask Delphi, but carry over more substantially.[18]

Against the backdrop of the above example types of prospective AI moral enhancement, what emerges is a kind of dilemma. On the one hand, prospective AI moral enhancement that consists in recommending to the user what is consistent with her pre-existing reported (or revealed) values is not clearly genuine *enhancement.* On the other hand, attempting to overcome this limitation by making prospective AI more robustly prescriptive (as in the case of "Ask Delphi") then runs in to a challenging version of the alignment problem and also faces (a version of) Archard's autonomy problem.

[15] Perhaps, through some combination of cognitively enhanced moral philosophers working in collaboration with advanced machine learning algorithms, we discover some optimal set of moral values to 'program in' to a moral enhancement AI, such that the prescriptive advice that the AI would give would be optimal (never mind how), thus in line with our objectives.

[16] See also Bang et al. (2023).

[17] https://delphi.allenai.org

[18] An interesting and broadly related earlier approach is developed by Klincewicz (2016), which uses an artificial moral reasoning engine to present moral arguments that are based in first-order normative theories (e.g., such as utilitarianism and Kantianism). In so far as the engine would be showing what consistent reasoning looks like from premises of such theory that one already accepts, the view might seem akin to the Giubilini-Savulescu coherence maximising approach. However, Klincewicz's moral reasoning engine is designed to play a stronger "normative" role, by potentially persuading users to accept its arguments conclusions. For instance, as Klincewicz says, "[…] it can, if prompted to do so, give answers to first-order normative questions, such as "should I report this to the authorities?" with a definite "yes" or "no" and then also provide reasons in support of that answer." In some respects, this latter characteristic of Klinewicz's proposal shares prescriptive commonalities with, e.g., "Ask Delphi".

With the above concerns in mind, consider a proposal that has the basic kind of structure that could in principle at least navigate the above dilemma, and which will be our working focus (in connection with the possibility of AI-enhanced moral expertise) in §§3-4. This is Francisco Lara and Jan Deckers (2019)'s proposal for moral enhancement via *Socratic AI*, sometimes just called SocrAI (Lara 2021).

Unlike AI that recommends choices that fit with an inputted hierarchy of values (or, like strong prescriptive AI, which simply prescribes recommendations in accordance with predefined values), the Lara and Deckers model focuses on how AI can teach improve our *moral reasoning*. With Socratic AI, the AI (e.g., imagine running an LLM such as a variant of ChatGPT) asks the agent questions that aim e.g., to help to clarify existing beliefs and values and encourage the agent to uncover option space they may have missed thus far. This type of activity, Lara and Deckers suspect, will not only help the agent exercise and hone their moral reasoning skills but also *motivate* agents to make choices that align with what they think is right, after reasoned reflection.[19] Notice that Socratic AI avoids the second horn of the dilemma in that (without requiring built in values) it avoids the kind of alignment problem that faces "Ask Delphi"; likewise, by not *recommending* particular courses of action, it sidesteps (also on the second horn) Archard's autonomy problem; Socratic AI is not encouraging you to *accept propositional moral content as true*. At the same time, there is some scope for dodging the first horn: to see why, compare Socratic AI versus, e.g., Giubilini/Savulescu's coherence maximising moral AI, in the case of someone who *begins with* (to simplify things here) bad values. Where will they end up, after interacting with these AIs, respectively? Whereas coherence-maximising moral AI simply recommends behaviour consistent with the bad values, Socratic AI prompts critical reflection that might potentially result in the *modification* of those values, including through improved moral reasoning skills that the Socratic AI aims to facilitate. In this way, we can see how the imagined individual might be genuinely morally better off via interaction with Socratic AI.

3. Is AI-Assisted Moral Enhancement Compatible with Genuine Moral Expertise?

Let's briefly take stock. We've seen that Socratic AI offers us at least one *prima facie* promising approach to AI-based moral enhancement, in that it offers a way we can rely on information (even if via sheer deference to prescriptive advice) from an AI assistant in a way that both (i) looks capable of leading to genuine moral improvement in a user; and (ii) does so without inviting the kinds of objections applicable to, e.g., strong prescriptive moral AIs.

Here is not the place to take any kind of definitive stance on whether Socratic AI *should* be pursued, or whether it is even successful (all things considered) as a moral enhancement strategy. Rather, the fact that we've seen (§2) that Socratic AI holds some particular promise compared to other would-be AI-based moral enhancement strategies makes Socratic AI a particularly fruitful form of AI-based moral enhancement.

The question we turn to now is whether *dependence on Socratic AI is compatible with possessing moral expertise?* The question of who the moral experts are, and in connection with wider

[19] For a recent and related kind of proposal, see Volkman and Gabriels (2023).

discussions of the value of moral expertise, make the possibility that technology such as Socratic AI could help generate moral expertise in users especially salient. It also is suggestive of the idea of a kind of 'democratization' of moral expertise: if Socratic AI were widely available then, ceteris paribus, so will the capacity to attain moral expertise.

This section will attempt to bring this guiding question of whether dependence on Socratic AI is compatible with attaining moral expertise under intellectual control, by identifying some key desiderata that any kind of AI-based moral enhancement would do well to satisfy if that enhancement should be thought to give rise to moral expertise; §4 then, with reference to these desiderata, looks at how Socratic AI holds up.

Before getting into these desiderata, though, I want to first make a few simplifying assumptions that will need made to get the rest of the discussion off the ground. Is there really moral expertise? Some philosophers have thought not – pointing to reasons ranging from the fact that, as Ryle noted, it's not obvious that the difference between right and wrong is something capable of being 'forgotten', to the fact that (as C.D. Broad (1952) put it), moral philosophers aren't in the business of telling people what to do[20] (even if experts in general are in such a position, in their relevant domains). Others, such as Driver (2013) and J.S. Gordon (2023), think that arguments that have dismissed the possibility of moral expertise have been too quick.

Let's assume there is such a thing as moral expertise – viz., expertise in the domain of morality.[21] What would be the *nature* of such expertise be? What would it plausibly involve, on the assumption it is a real phenomenon humans can aspire to?

While it might be tempting to venture few plausible necessary conditions on moral expertise, I want to try to avoid doing so; this is because conditions sufficient for moral expertise might not also be necessary. Regardless of whether we think (as per Weinstein 1993) that epistemic expertise (roughly: expertise consisting in 'providing strong justifications for a range of proposition in a domain' and performative expertise (roughly: expertise consisting in 'the capacity to perform a skill well according to the rules and virtues of a practice' mark out genuinely different *types* of expertise (or whether they are beset understood as different realisations of the same type), we have empirical evidence that suggests that experts who perform well might not always be knowledgeable in the way that would seem sufficient in some contexts of being an expert. As Matt Stichter (2015, 113) puts it: "Even when experts are able to articulate an explanation, the explanations are often inconsistent with the observed behaviour of the experts."

Rather than to assume then that if moral expertise exists it would involve meeting any necessary conditions, let's – for our purposes here – simply take note of a cluster of *dimensions* that are widely taken to track expertise. On the assumption that expertise (like cognate notions of skill and know-how) is a matter of degree, we might then think that an individual is a better

[20] As Broad (1952) puts it, "It is no part of the professional business of moral philosophers to tell people what they ought or ought not to do. . . . Moral philosophers, as such, have no special information not available to the general public, about what is right and what is wrong; nor have they any call to undertake those hortatory functions which are so adequately performed by clergymen, politicians, leader-writers".

[21] While some discussions of expertise relative expertise to a skill as well as a domain (where we might think of a skill as a trait manifested within a wider domain – see Stichter (2015), for simplicity I'll talk of expertise in connection with domains.

candidate for moral expertise the more of these dimensions they meet (without taking the failure of any particular dimension to be disqualifying).

These simplifying assumptions made, what are some of the dimensions of moral expertise? Literature on expertise generally (and moral expertise specifically) suggests a varying range. Here I want to note *six key dimensions* which different philosophers have identified, and in §4 we'll use these as provisional criteria to return in a more organised way to our question about Socratic AI and moral expertise.

One straightforward metric on which expertise is attributed is *knowledgeableness* (e.g., (Goldman 2001, 91); we expect an expert ornithologist to know a lot about birds. In a similar vein, a moral expert should be knowledgeable about the domain of morality.[22]

A separate metric however tracks *performance*: we expect, e.g., an expert in the law to not merely engage knowledgeably in reflection on legal facts, but to *manifest* this knowledge in their performance (Weinstein 1993) – this might involve following laws, pointing out when a law is broken and what counts as the breaking of particular laws, etc. In the moral case, could envision – analogously – the attribution of moral expertise tracking something like *manifesting* moral knowledge in action.

A third expertise metric acknowledged in the literature is *automaticity*. As Stichter (2015) puts it, "Experts do not need to devote much conscious attention to what they are doing, and this lack of conscious attention does not lead to any reduction in their performance" (2015, 64). Translated to specifically the moral case: we might expect a moral expert might, as Aristotle would expect of the morally virtuous, to do the right thing without prior deliberation on, e.g., moral principles. This point lines up likewise with thinking about expertise due to Hubert Dreyfus, who denies that expertise involves the conscious following of rules (moral or otherwise).[23]

A fourth dimension that's been associated with expertise is *rational autonomy*. As Finnur Dellsén (2020, 358) argues in a recent paper: "Experts should make up their own minds about issues that fall within their domain of expertise, as opposed to following the opinions of their fellow experts"; moral expertise, by this metric, would be associated with forming moral views in a way that is not merely deferential, including, by deference to other experts.

A fifth metric associated with expertise is *principle unification* – as defended in work on expertise and virtue by Julia Annas (1995). As Annas sees it, domains that admit of skill and expertise are governed by various unifying principles, and the expert must have a grasp of not only part of the field. A chess expert grasps principles governing good openings playing black and playing white, openings and end games, etc. and not just some principles lining up with a 'part' of chess strategy. By parity of reasoning, we might associate moral expertise with having more than a mere partial grasp of moral principles, viz., more than just a grasp of *some* principles.

A sixth and final metric of expertise, discussed in various ways by Alvin Goldman (2018) and Christian Quast (2018), though cf., Michel Croce (2019), is *helpfulness*. As Goldman puts it, in attributing expertise, we consider not just the expert's own traits in isolation from their community, but also 'what experts can *do* for laypersons by means of their special knowledge or skill'. Part of genuine expertise, for Goldman, is having 'the capacity to help others (especially laypersons) solve a variety of problems' in the domain in which one is an expert. A medical

[22] This knowledge might take different forms (know-how, propositional knowledge, occurrent, tacit, etc.)

[23] See also, relatedly, discussion of moral expertise and fluency due to Railton (2008).

expert can help a patient treat an illness; and by parity of reasoning, a moral expert by this metric will be positioned to be morally *helpful* – viz., to (perhaps) help others think through moral problems well.

4. Socratic AI and Moral Expertise, revisited

What we've gained now from the previous section is that (i) on the simplifying assumption that there is such a thing as moral expertise, we can expect that (ii) the extent to which one attains moral expertise could be reasonably expected to track the six different recognised expertise metrics detailed in §3, and we saw roughly what those metrics, in the case of moral expertise specifically, might look like if satisficed.

Against that background, let's return now to Socratic AI. By way of reminder, the question under investigation here isn't whether Socratic AI *itself* might aspire to expertise. That question takes us well beyond what I'm aiming to cover here, as it raises questions (including questions of increasing interest in bioethics) about whether artificial agents might be literal bearers of agency properties[24], like knowledge and expertise. (See, however, Rodríguez-López and Rueda 2023 for an recent case defending at least some kinds of AI as bona fide moral experts).

Rather, let's consider the extent to which dependence on Socratic AI-based moral enhancement might track the possession of genuine moral expertise, taking each of the six metrics in turn.

First, consider *knowledgeableness*. There is well-known scepticism in the literature on moral deference about whether one can gain moral *knowledge* via testimony[25], and this is a serious strike against the thought that one could gain moral expertise (at least by this metric) simply through consistent reliance on even the best kind of strong prescriptive AI (e.g., such as "Ask Delphi"). The concern is, roughly, that there is something distinctive about, e.g., moral and aesthetic beliefs, which is that we are in the market for knowledge of such beliefs only by in some sense appreciating for ourselves why they are true when they are. In so far as sheer deference allows one to believe a proposition in the absence of such appreciation for its grounds, sheer deference isn't a route to moral (and aesthetic) knowledge, or so the argument goes.

Interestingly for our purposes, notice how this line of argument does not carry over from reliance on strong prescriptive moral AI to reliance on Socratic AI. Here an analogy between knowledge gained through psychotherapy and Socratic AI is instructive. Let's take as a starting point the premise that cognitive behavioural therapy (CBT) is a potential avenue for gaining self-knowledge – in so far as it facilitates change in one's beliefs about oneself in par through targeting and changing irrational aspects of an agent's thinking about herself. Consider now that interacting with Socratic AI appears to be akin to interacting with a very narrowly focused cognitive behavioural therapist, insofar as both cognitive behavioural therapy (CBT) and Socratic AI work to find inconsistent and irrational aspects of an agent's thinking (e.g., Overholser 2010). However, there's at least one way in which Socratic AI is likely to be more effective than a CBT practitioner. While the cognitive behavioural therapist has to work to set aside their own values, the Socratic AI simply doesn't come with particular values or choices of values—Lara and

[24] For recent discussion on this point, see, e.g., Cervantes et al. (2020) and Bryson, Kime, and Zürich (2011).

[25] See, e.g., McGrath (2009) and Hills (2009).

Deckers (p.281-2) propose the system have "no previous lists or systems of values from which to improve the morality of the agent", with algorithms designed to avoid the machine being biased towards any particular values or normative ethical theories. Instead "through the constant interaction between the agent and the system, the possibility that the agent's values would be changed through their dialogue with the machine is increased" (Lara and Deckers 2019, 281). If (non-prescriptive) CBT is a route to self-knowledge (even if not by prescribing an agent accept beliefs on sheer deference), then we have reason to think Socratic AI would likewise be an analogous kind of route to knowledge in the moral domain.

Suppose then that a user of AI gains moral knowledge through her dependence on Socratic AI. Once this much is granted, is there any barrier to the user of Socratic AI manifesting moral knowledge acquired in action? A critic here might suggest that Socratic AI, even if capable of inculcating moral knowledge (in a way broadly analogous to the way CBT might do so) the knowledge is not going to be as situation specific or actionable as, e.g., the kind of moral *information* that we could expect form a strong prescriptive moral AI. Take for example an imagined case where you are in doubt about whether to give a particular friend who has fallen short in the past a second chance with a high-responsibility task. Whereas a strong prescriptive moral AI might simply communicate action-guiding information here, Socratic AI will not do so. Is this much a genuine barrier to a user of Socratic AI attaining moral expertise along as captured by a performance metric on expertise? There's a good case here for pressing back with some optimism. Imagine here a 'good case' where one's moral values improve over time through the use of Socratic AI, through reflective and thoughtful interactions with the AI, which prompt moral belief revision in a way where rationally supported moral knowledge (e.g., perhaps of certain moral principles) is attained. *Manifesting* moral knowledge in one's actions might very well manifest known *principles* rather than merely known (situation specific) moral information one might have in a particular situation. For example, your repaying a loan might manifest your general knowledge that loans should be repaid, rather than any specific knowledge about whether you should repay a particular loan on occasion. Once this point is appreciated, though, initial reservations for doubting that Socratic AI interactions might support expert moral action seems overstated.

What about *automaticity?* This might look, initially, like the most serious disanalogy between what we'd expect of a domain expert and what we'd envision in the case of a user (in the moral domain) of Socratic AI. Here is perhaps the strongest form of the challenge: following Stichter, 'Experts do not need to devote much conscious attention to what they are doing' (2015, 64). A user of Socratic AI, however, is consciously engaging with the AI, depending on its responses, in a way we might conceptualise as a kind of intermediate and conscious step, a 'thinking and an acting' that's incongruous with the automaticity of an expert. This is a fair point of criticism. However, there is perhaps an equally compelling line of reply – one that is concessionary in that it simply grants that the *consultation* of Socratic AI is a conscious activity, one involving conscious attention, which is as such an activity that is at odds with automaticity. However, as this line of thought goes, we should distinguish between the learning phase (when one is interacting with Socratic AI), and whatever moral expertise would be the *result* of such learning. If we think of Socratic AI as a kind of 'scaffolding', whereby one (through interaction with the AI) learns morally over time, then we can grant the critic that there is no automaticity during the learning stage. *But*, as the thought goes, this is just what we should expect of learning in other domains, where conscious attention to one's action during learning *precedes* expertise. In the case of moral expertise attained through Socratic AI, then, the thought would be that (in connection with automaticity) the conscious attention to *learning* is not disanalogous with what we find in other domains of expertise; and further, following a period of moral leaning through Socratic AI, we can expect one's moral knowledge-manifesting action would be similarly automatic.

What about *rational autonomy?* In so far as (as per Dellsén 2020) experts as such should "make up their own minds", moral expertise not only would preclude, e.g., either deference *or* agnosticism, but it would also be such that it could be improved through gaining skills to improve at making up one's mind in a moral matter in a rational way. With this in mind, not only would moral enhancement via Socratic AI avoid the promotion of moral deference (as already noted) but it functions so as to improve one morally specifically by empowering the user's rational autonomy via improved moral reasoning skills. On this point, Lara and Deckers give several example illustrations of how Socratic AI users might become better moral reasoners:

- By being assisted in anticipating and explaining the likely consequences of particular choices.
- By learning about how aspects of our environment impact decision-making.
- By being made aware of how human biology impacts decision-making.
- By having ambiguous language use pointed out.
- By having clarity highlighted and encouraged.
- By learning about empirical support for particular beliefs (and learning about the *lack* of empirical support for particular beliefs). (Lara and Deckers 2020)

The only viable line of rejoinder I see in the case of rational autonomy, on behalf of the critic of the prospects of gaining moral expertise through Socratic AI, would hold that one's rational autonomy might be in some way diminished by her *epistemic dependence* on the Socratic AI itself. This objection overgeneralises, however. If relying on dialogue to facilitate moral reasoning is rational autonomy-undermining, then presumably it will likewise be autonomy undermining in more traditional forms of education that avoid indoctrination by simply aiming to facilitate critical thinking skills. Put another way, the bare dependence on Socratic AI isn't plausible rational autonomy undermining unless we accept, implausibly, trivialise the undermining of rational autonomy so as to grant that it occurs in more standard cases of acquiring critical thinking skills in learning.

What about Annas's 'principle unification' metric of expertise? On the assumption that domains that admit of expertise are governed by constitutive principles, should we think non-prescriptive character of Socratic AI would somehow prevent one from coming to learn such principles? On this point, it's worth considering how this argument might be turned completely around. If we take seriously Annas's suggestion that the expert has a 'grasp' of the relevant principles in the domain, then there is scope to think that prescriptive AI not only isn't needed to gain knowledge of moral principles, but that it would not be a candidate for disseminating such knowledge. Consider here, as is suggested by the moral deference literature, that coming to believe prescribed principles is not a route to grasping (of the sort we might identify with *understanding*). If that is right, then rather than to view Socratic dialogue of the sort facilitated by Socratic AI as not sufficient for generating knowledge of moral principles, we should think that if one can gain knowledge of such principles at all, the way in which Socratic dialogue might facilitate grasping – a point I've argued for elsewhere[26] – suggests Socratic AI might be particularly well placed to help a user gain such knowledge.

Let's consider now the expertise metric of 'helpfulness', defended variously by Goldman and Quast. One might think – imagining a critical stance here – that a user is prevented from giving useful moral guidance if she is depending herself on external scaffolding. As the thought might go, the credit for the helpfulness goes to the Socratic AI, not to the user. There are two lines of response here. First, the general principle underlying the reasoning overgeneralises. We don't say that expert air traffic controllers, for instance, lack the kind of 'helpfulness' apposite to expertise given that their helpfulness is predicated upon their dependence on the computers they require to track airplane patterns in real time. Second, and perhaps more importantly, even if we granted that dependence on something external to one is in some way at odds with the kind of helpfulness to laypersons befitting of expertise, the objection here would be at most applicable to initial learning stages with Socratic AI. Recall the respond to the anticipated objection from automaticity: after prolonged use of Socratic AI, one might (as with other learning processes) be well positioned to manifest her moral knowledge through the offering of guidance. If the premise that one can gain moral knowledge through one's interactions with Socratic AI over time is granted, then so should be the associated idea that downstream manifestation of that knowledge could be put to the service of assisting others in a way we'd expect an expert in any domain to be in a position to do.

Summing up, then, it looks like we have every reason to be optimistic Socratic AI is a genuine route to moral expertise.[27] And even more than that, on the assumption that expertise is a matter of degree, and that the six metrics identified are plausible cluster critera for identifying experts (even if not individually necessary conditions), we've got cause to be optimistic that the kind of moral enhancement that would be facilitated through Socratic AI is capable of leading to a significant level of moral expertise – that is, moral expertise that lines up with an array of the most typical markers of expertise we should expect of in any domain.

This result has an interesting bearing on the connection between moral enhancement and moral expertise, which is that – to put it simply – the very idea of enhanced expertise in the moral domain is about as viable as the idea of moral enhancement through AI more generally. That is, in so far as reliance on Socratic AI (even if not necessarily on other versions of AI-based moral enhancement) is a genuine form of moral enhancement (a conclusion that looked promising in §2), we have good cause to think that there is no barrier between the enhancement one would attain and the inculcation of genuine moral expertise.

This is a welcome result at least from the perspective where ceteris paribus the democratisation of expertise is valued. Whereas in some domains, expertise might be available only to certain privileged or elite, we've got good reason to think that in so far as Socratic AI could be widely available – as we already expect it might be given the recent emergence of LLMs when can be accessible (e.g., much like OpenAI's ChatGPT from a mobile phone -- a more egalitarian distribution of moral expertise is not implausible, at least for those who have the desire (as with therapy) to put in the reflective work.

[26] See on this point Gordon (2016).

[27] It's worth noting that I take the optimism here to simply apply to Socratic AI as one, among potentially many, routes to moral expertise. For instance, it might also be that, entirely independently of the benefits one can attain through Socratic AI, engaging with exemplars or role models can have important moral benefit. Thanks to a referee for raising this point.

5. Concluding Remarks

Artificial intelligence offers one of the most promising new routes to moral enhancement. Separately, the question of *moral expertise* has remained an important one at the forefront of ethics and its intersection with social epistemology. The aim here has been to bring these debates together, to show that a viable form of moral enhancement via artificial intelligence is at the same time a viable route to moral expertise. §1 introduced some of the guiding themes of these debates, §2 distinguished between several varieties of AI-based moral enhancement, and showed how Socratic AI has some potential advantages over other proposals in so far as it offers a genuine route to moral enhancement. §§3-4 then evaluated whether Socratic AI might be in the market not only for supporting moral enhancement but full-blown moral expertise. §3 outlined six metrics associated with expertise in a given domain generally, with discussion of what this would look like in the moral domain specifically; §4 then considered how users depending on Socratic AI measure up to these metrics, concluding with some optimism that in so far as Socratic AI is a route to moral enhancement, it is likewise a route to moral expertise. What follows more generally is welcome result about the potential democratisation of moral expertise.

References

Annas, Julia. 1995. 'Virtue as a Skill'. *International Journal of Philosophical Studies* 3 (2): 227–43.

Archard, David. 2011. 'Why Moral Philosophers Are Not and Should Not Be Moral Experts'. *Bioethics* 25 (3): 119–27.

Bang, Yejin, Nayeon Lee, Tiezheng Yu, Leila Khalatbari, Yan Xu, Samuel Cahyawijaya, Dan Su, et al. 2023. 'Towards Answering Open-Ended Ethical Quandary Questions'. arXiv. https://doi.org/10.48550/arXiv.2205.05989.

Berry, Ed D. J., Richard J. Allen, Mark Mon-Williams, and Amanda H. Waterman. 2019. 'Cognitive Offloading: Structuring the Environment to Improve Children's Working Memory Task Performance'. *Cognitive Science* 43 (8): e12770. https://doi.org/10.1111/cogs.12770.

Bostrom, Nick, and Anders Sandberg. 2009. 'Cognitive Enhancement: Methods, Ethics, Regulatory Challenges'. *Science and Engineering Ethics* 15 (3): 311–41.

Broad, C. D. 1952. 'Ethics and the History of Philosophy: Selected Essays'.

Bryson, Joanna J., Philip P. Kime, and C. Zürich. 2011. 'Just an Artifact: Why Machines Are Perceived as Moral Agents'. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 22:1641. Citeseer.

Carter, J. Adam, and Duncan Pritchard. 2019. 'The Epistemology of Cognitive Enhancement'. In *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, 44:220–42. Oxford University Press US.

Cervantes, José-Antonio, Sonia López, Luis-Felipe Rodríguez, Salvador Cervantes, Francisco Cervantes, and Félix Ramos. 2020. 'Artificial Moral Agents: A Survey of the Current Status'. *Science and Engineering Ethics* 26: 501–32.

Cholbi, Michael. 2007. 'Moral Expertise and the Credentials Problem'. *Ethical Theory and Moral Practice* 10: 323–34.

Christian, Brian. 2020. *The Alignment Problem: Machine Learning and Human Values*. WW Norton & Company.

Constantinescu, Mihaela, Constantin Vic\ua, Radu Uszkai, and Cristina Voinea. 2022. 'Blame It on the AI? On the Moral Responsibility of Artificial Moral Advisors'. *Philosophy and Technology* 35 (2): 1–26. https://doi.org/10.1007/s13347-022-00529-z.

Croce, Michel. 2019. 'Objective Expertise and Functionalist Constraints'. *Social Epistemology Review and Reply Collective* 8 (5): 25–35.

Dellsén, Finnur. 2020. 'The Epistemic Value of Expert Autonomy'. *Philosophy and Phenomenological Research* 100 (2): 344–61.

Driver, Julia. 2013. 'MORAL EXPERTISE: JUDGMENT, PRACTICE, AND ANALYSIS'. *Social Philosophy and Policy* 30 (1–2): 280–96. https://doi.org/10.1017/S0265052513000137.

Earp, Brian D. 2018. 'Psychedelic Moral Enhancement'. *Royal Institute of Philosophy Supplements* 83 (October): 415–39. https://doi.org/10.1017/S1358246118000474.

Earp, Brian D., Anders Sandberg, and Julian Savulescu. 2012. 'Natural Selection, Childrearing, and the Ethics of Marriage (and Divorce): Building a Case for the Neuroenhancement of Human Relationships'. *Philosophy & Technology* 25 (4): 561–87. https://doi.org/10.1007/s13347-012-0081-8.

Gilbert, Sam J., Arabella Bird, Jason M. Carpenter, Stephen M. Fleming, Chhavi Sachdeva, and Pei-Chun Tsai. 2020. 'Optimal Use of Reminders: Metacognition, Effort, and Cognitive Offloading.' *Journal of Experimental Psychology: General* 149 (3): 501.

Giubilini, Alberto, and Julian Savulescu. 2018. 'The Artificial Moral Advisor. The Ideal Observer Meets Artificial Intelligence'. *Philosophy and Technology* 31 (2): 169–88. https://doi.org/10.1007/s13347-017-0285-z.

Goldberg, Sanford. 2009. 'Experts, Semantic and Epistemic'. *Noûs* 43 (4): 581–98. https://doi.org/10.1111/j.1468-0068.2009.00720.x.

Goldman, Alvin I. 2001. 'Experts: Which Ones Should You Trust?' *Philosophy and Phenomenological Research* 63 (1): 85–110. https://doi.org/10.1111/j.1933-1592.2001.tb00093.x.

———. 2018. 'Expertise'. *Topoi* 37 (1): 3–10.

Gordon, Emma. 2016. 'Social Epistemology and the Acquisition of Understanding'. In *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, 293–317. Routledge.

Gordon, Emma C. 2022a. 'Cognitive Enhancement and Authenticity: Moving beyond the Impasse'. *Medicine, Health Care and Philosophy*, 1–8.

———. 2022b. *Human Enhancement and Well-Being: A Case for Optimism*. Taylor & Francis.

Gordon, Emma C., and Rebecca J. Willis. 2023. 'Pharmacological Cognitive Enhancement and the Value of Achievements: An Intervention'. *Bioethics* 37 (2): 130–34.

Gordon, John-Stewart. 2023. 'Moral Expertise Revisited'. *Bioethics* 37 (6): 533–42. https://doi.org/10.1111/bioe.13172.

Grinschgl, Sandra, Hauke S. Meyerhoff, and Frank Papenmeier. 2020. 'Interface and Interaction Design: How Mobile Touch Devices Foster Cognitive Offloading'. *Computers in Human Behavior* 108 (July): 106317. https://doi.org/10.1016/j.chb.2020.106317.

Gupta, Kristina. 2012. 'Protecting Sexual Diversity: Rethinking the Use of Neurotechnological Interventions to Alter Sexuality'. *AJOB Neuroscience* 3 (3): 24–28.

Hannon, Michael. 2021. 'Disagreement or Badmouthing? The Role of Expressive Discourse in Politics'. In *Political Epistemology*, edited by Elizabeth Edenberg and Michael Hannon. Oxford: Oxford University Press.

Hardcastle, Valerie Gray. 2018. 'Lone Wolf Terrorists and the Impotence of Moral Enhancement'. *Royal Institute of Philosophy Supplements* 83: 271–91.

Harris, John. 2011. 'Moral Enhancement and Freedom'. *Bioethics* 25 (2): 102–11.

Hills, Alison. 2009. 'Moral Testimony and Moral Epistemology'. *Ethics* 120 (1): 94–127.

Hurlemann, René, Alexandra Patin, Oezguer A. Onur, Michael X. Cohen, Tobias Baumgartner, Sarah Metzler, Isabel Dziobek, Juergen Gallinat, Michael Wagner, and Wolfgang Maier.

2010. 'Oxytocin Enhances Amygdala-Dependent, Socially Reinforced Learning and Emotional Empathy in Humans'. *Journal of Neuroscience* 30 (14): 4999–5007.

Jiang, Liwei, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, et al. 2022. 'Can Machines Learn Morality? The Delphi Experiment'. arXiv. https://doi.org/10.48550/arXiv.2110.07574.

Jotterand, Fabrice, and Susan B. Levin. 2019. 'Moral Deficits, Moral Motivation and the Feasibility of Moral Bioenhancement'. *Topoi* 38: 63–71.

Kass, Leon. 2002. *Life, Liberty and the Defense of Dignity: The Challenge for Bioethics*. Encounter books.

Klincewicz, Michał 2016. 'Artificial Intelligence as a Means to Moral Enhancement', *Studies in Logic, Grammar and Rhetoric* 48 (1):171-187.

Lara, Francisco. 2021. 'Why a Virtual Assistant for Moral Enhancement When We Could Have a Socrates?' *Science and Engineering Ethics* 27 (4): 42. https://doi.org/10.1007/s11948-021-00318-5.

Lara, Francisco, and Jan Deckers. 2020. 'Artificial Intelligence as a Socratic Assistant for Moral Enhancement'. *Neuroethics* 13 (3): 275–87. https://doi.org/10.1007/s12152-019-09401-y.

Liao, S. Matthew. 2011. 'Parental Love Drugs: Some Ethical Considerations'. *Bioethics* 25 (9): 489–94. https://doi.org/10.1111/j.1467-8519.2009.01796.x.

Liu, Yuxin, Adam Moore, Jamie Webb, and Shannon Vallor. 2022. 'Artificial Moral Advisors: A New Perspective from Moral Psychology'. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 436–45. AIES '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3514094.3534139.

Maslen, Hannah, Filippo Santoni de Sio, and Nadira Faber. 2015. 'With Cognitive Enhancement Comes Great Responsibility?' *Responsible Innovation 2: Concepts, Approaches, and Applications*, 121–38.

McGrath, Sarah. 2009. 'The Puzzle of Pure Moral Deference'. *Philosophical Perspectives* 23: 321–44.

Melo-Martín, Inmaculada de. 2018. 'The Trouble With Moral Enhancement'. *Royal Institute of Philosophy Supplement* 83: 19–33. https://doi.org/10.1017/s1358246118000279.

Mikolajczak, Moïra, Nicolas Pinon, Anthony Lane, Philippe de Timary, and Olivier Luminet. 2010. 'Oxytocin Not Only Increases Trust When Money Is at Stake, but Also When Confidential Information Is in the Balance'. *Biological Psychology* 85 (1): 182–84. https://doi.org/10.1016/j.biopsycho.2010.05.010.

Nisbett, Richard E., and Timothy D. Wilson. 1977. 'Telling More than We Can Know: Verbal Reports on Mental Processes.' *Psychological Review* 84 (3): 231.

Nyholm, Sven. 2015. 'Love Troubles: Human Attachment and Biomedical Enhancements'. *Journal of Applied Philosophy* 32 (2): 190–202.

Overholser, James C. 2010. 'Psychotherapy According to the Socratic Method: Integrating Ancient Philosophy with Contemporary Cognitive Therapy'. *Journal of Cognitive Psychotherapy* 24 (4): 354–63.

Persson, Ingmar, and Julian Savulescu. 2008. 'The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity'. *Journal of Applied Philosophy* 25 (3): 162–77. https://doi.org/10.1111/j.1468-5930.2008.00410.x.

———. 2012. *Unfit for the Future: The Need for Moral Enhancement*. OUP Oxford.

Pokorny, Thomas, Katrin H Preller, Michael Kometer, Isabel Dziobek, and Franz X Vollenweider. 2017. 'Effect of Psilocybin on Empathy and Moral Decision-Making'. *International Journal of Neuropsychopharmacology* 20 (9): 747–57. https://doi.org/10.1093/ijnp/pyx047.

Quast, Christian. 2018. 'Expertise: A Practical Explication'. *Topoi* 37 (1): 11–27. https://doi.org/10.1007/s11245-016-9411-2.

Railton, Peter. 2009. 'Practical competence and fluent agency', In David Sobel & Steven Wall (eds.), *Reasons for Action.* Cambridge University Press. pp. 81--115.

Riaz, Amber. 2020. ‘How to Identify Moral Experts’. *The Journal of Ethics* 25 (1): 123–36. https://doi.org/10.1007/s10892-020-09338-y.

Rodríguez-López, Blanca, and Jon Rueda. 2023. ‘Artificial Moral Experts: Asking for Ethical Advice to Artificial Intelligent Assistants’. *AI and Ethics*, 1–9.

Sandel, Michael J. 2007. *The Case against Perfection: Ethics in the Age of Genetic Engineering*. Harvard university press.

Schwitzgebel, Eric. 2008. ‘The Unreliability of Naive Introspection’. *Philosophical Review* 117 (2): 245–73.

Stichter, Matt. 2015. ‘Philosophical and Psychological Accounts of Expertise and Experts’. *HUMANA.MENTE Journal of Philosophical Studies* 8 (28): 105–28. https://www.humanamente.eu/index.php/HM/article/view/83.

Tassella, Marco, Rémy Chaput, and Mathieu Guillermin. 2023. ‘Artificial Moral Advisors: Enhancing Human Ethical Decision-Making’. In , 1. IEEE. https://doi.org/10.1109/ETHICS57328.2023.10155026.

Volkman, Richard, and Katleen Gabriels. 2023. ‘AI Moral Enhancement: Upgrading the Socio-Technical System of Moral Engagement’. *Science and Engineering Ethics* 29 (2): 11. https://doi.org/10.1007/s11948-023-00428-2.

Wang, Ju. 2021. ‘Cognitive Enhancement and the Value of Cognitive Achievement’. *Journal of Applied Philosophy* 38 (1): 121–35.

Weinstein, Bruce D. 1993. ‘What Is an Expert?’ *Theoretical Medicine* 14 (1): 57–73. https://doi.org/10.1007/BF00993988.

Wilson, Clare. 2020. ‘New Apple Watch Monitors Blood Oxygen–Is That Useful?’ Elsevier.

Yudkowsky, Eliezer. 2016. ‘The AI Alignment Problem: Why It Is Hard, and Where to Start’. *Symbolic Systems Distinguished Speaker* 4.