AI and the Pedagogy of Ignorance[1]

Brian Hamilton
Florida Southern College

ABSTRACT: As ethics teachers consider whether and how to incorporate new forms of generative artificial intelligence into their classrooms, it is important to revisit the question of what exactly it is that we aim to teach. We ought to use these technologies only if we know that they help us do better what we claim to do. To explore this question, this paper reconsiders the founding provocation of Western philosophy: Socrates's claim that true wisdom consists in knowing that one is not wise. I argue that Socrates meant that claim earnestly, and that it is a legitimate and practicable aim of ethics education. I contrast Socrates's view with that of Protagoras, the famous Greek sophist, who proposes instead that ethics education aims at teaching sound deliberation. On either view, I argue, tools like ChatGPT have little to offer ethics teachers—and on the Socratic view, nothing at all.

YOU KNOW THE STORY: a certain Chaerephon, an impetuous but well-liked Athenian democrat, hikes up to Delphi and asks the priestess of the temple there whether there is anyone wiser than his dear friend Socrates.[2] She says there is not. This comes as quite a shock to Socrates, who doesn't think himself wise at all. He decides there must be some riddle here (indeed, with the Pythia there always was) and sets out to untangle it. He goes first to some man-about-town with an impressive reputation and begins to ask him questions, only to discover that he's not nearly as wise as people think he is—not nearly as wise as *he* thinks he is. Socrates catches his first glimpse of what the riddle might mean: at least he is not under the delusion that he is wise when in fact he is not. Is that a kind of wisdom in itself? Socrates continues, person by person, group by group, through the citizens of Athens, talking to anyone who is supposed to know anything. He talks to people in politics and the arts and the trades. Everywhere he finds the same thing. Those most respected for their wisdom knew the least, while those least respected knew the most. Foolishly, perhaps—we'll have to consider how foolish it was—Socrates became obsessed with trying to *prove* to people that they weren't as wise as they thought they were. This, he thought, was the answer to the riddle and a direct charge from the God. Socrates decided he must really

be the wisest, or at least he had the attitude of someone who was truly wise, precisely because he knew that he was *not* wise and that in fact what usually passes for wisdom was worthless. His singular goal in life, his lonely battle post, would be to help others understand this fact. He succeeded, however, mainly in making people hate him. According to his enemies, he was undermining the fragile constitution of Athens by making a mockery of its protectors. Not content to ostracize him—or rather, because Socrates definitively refused to leave town or let up—his compatriots finally executed him by majority vote.

This is easily the most famous story in the history of philosophy, and Socrates the most influential of philosophers. Yet the idea that wisdom consists in knowing that we are not wise remains a difficult one to take seriously, especially for teachers, committed as we usually are to the idea that we *do* know something and that our goal is to get our students to know that thing, too. So this idea, which the Socrates of the *Apology* presents as his one and only doctrine,[3] is received, if at all, only as a quick sermon about epistemic humility, a reminder not to get too cocky, perhaps a reason to assign confidence intervals to our claims, before we settle into the real work of making good arguments and getting things right.

What is it, exactly, that ethics teachers claim to teach? What knowledge or wisdom is it that we claim to hand on to our students? This is a very basic question, but an important one, especially in this new era of generative artificial intelligence that can speak, apparently, like us. We are under significant pressure to decide whether and how to incorporate tools like ChatGPT into our classrooms—or rather, how to respond to the fact that our students are bringing these tools into our classrooms whether we like it or not. Because we live in the age we do, many automatically suppose that the value of this technology is self-attesting: we *should* use this technology simply because this technology, with all its sparkling new capabilities, exists. But we should of course resist this presupposition. The central question to ask about any new technology is whether it helps or hinders us in achieving our ends. A *technē* is only as good as the good it helps to produce. But that principle requires us to be clear about what our ends are. It is obvious that a new species of wireless charger cannot help me teach ethics. It is not immediately obvious whether or not ChatGPT can. It is therefore important to revisit the basic question of what it is that we are teaching.

In this essay, I will set out two possible conceptions of the ends of ethics pedagogy—the Protagorean (after the Protagoras of Plato's dialogue by that name) and the Socratic—and I will consider whether and how ChatGPT and its like might be useful in each conception. Protagoras claimed to teach *euboulia*, sound deliberation, for the sake of the soul and the city. Socrates claimed to teach ignorance, for the same reasons.[4] In neither case, in my judgment, is generative AI very helpful, but in the Socratic view it is no use at all.

In the first section of this essay, I will say more about the Socratic embrace of ignorance. What did he really mean by claiming to know nothing? How seriously should we take those claims? Why was he so skeptical of teachers like

Protagoras? In the second and third sections, I will bring us back to the contemporary ethics classroom to consider how AI might or might not contribute to our pedagogical work.

The Moral Importance of Ignorance

It is easy to think, and many people have thought, that Socrates's professed ignorance was just a ruse. This is almost certainly what his enemies thought. Each time he examined someone new, Socrates admits, "the bystanders thought I myself was wise about the things I refuted in others."[5] They thought his claim to ignorance was meant to humiliate. And they might have had good reason to think so, since the sophists seem to have used public humiliation as a sales tactic—proving that if you paid up, you too could learn how to make fools of your opponents in public.[6] This is why Socrates makes sure to remind the jury that he has never been a professional teacher, has never charged for anything.[7]

But there's also a more sympathetic version of the idea that Socratic ignorance is a ruse, one that teachers still sometimes advocate. Claiming ignorance can be a good *pedagogical strategy*. It can be worthwhile, the thinking goes, to disclaim one's own expertise and to enter instead into the ignorance of one's students. Rather than standing over them as a sovereign dispenser of knowledge, it might be better to approach the material from the students' perspective, engaging with them in a process of inquiry starting from their own beliefs about the world—even if, in fact, we know where we want them to end up. Socrates might seem to be teaching in just this way. But he insists he wasn't. He insists he genuinely had no knowledge of the things he wanted to find out. He is not merely "setting aside his presuppositions" or "approaching with a beginner's mind" as a way of guiding his student in the right direction. Invariably, Socrates positions himself as the student and his interlocutor as the teacher, even if it turns out, as it always does, that his interlocutor has nothing to teach.

But if his ignorance is not a ruse, whether meant to humiliate or to instruct, what are we to make of it? There is something tragic, if not self-defeating, about a person who talks endlessly about the importance of knowledge, who seeks knowledge obsessively right up to the point of his death, who tells others that seeking such knowledge is the very point of life, and yet has none. Alexander Nehamas makes the point painfully clear when he asks whether Socrates dooms his own project to failure by insisting (apparently) that living well depends on being able to define what "living well" means, but also that arriving at such a definition is impossible.[8] Certainly from a pedagogical point of view, Socrates's case seems hopeless: a moral teacher who genuinely believes he knows nothing of virtue, who is not even sure that virtue can be taught.[9]

Gregory Vlastos, a towering figure in the interpretation of Socrates, tried to make sense of this problem by suggesting that Socrates thought of knowledge in two different ways.[10] Sometimes he thought of knowledge as *certain* knowledge, *infallible* knowledge—and this he disavowed earnestly and entirely. But

sometimes he thought of knowledge in a more modest way, as the conclusion of elenctic inquiry. Such conclusions can reasonably be considered knowledge, but they are far from infallible. It is always possible that the next inquiry will unearth a contradiction that has so far been overlooked. Socrates did, according to Vlastos, grant himself this second sort of knowledge. This second sort of knowledge is what he calls in the *Apology* properly human wisdom (*anthrōpinē sophia*); the first is what he calls a wisdom more than human.[11]

This explanation has the benefit of making Socrates infinitely more intelligible to us, to be sure, since it brings him comfortably into line with the epistemology of the modern natural sciences. The epistemology of the modern natural sciences takes the provisionality of all knowledge as its guiding principle: we can only ever make judgments *given the current evidence*, but we never claim a definitive judgment. Socrates might have different reasons for holding this principle—the elenchus is focused on the consistency of premises, while the natural sciences are focused on the possibility of countervailing data—but we intuitively recognize the importance of this sort of epistemological proviso. This explanation of Socratic ignorance is also fairly easy for us as teachers of ethics to assimilate. We already tell our students to stay open to new perspectives, to hold out the possibilities that their settled ethical judgments are mistaken. Test vigorously, form judgments, but do not settle into dogmatism. Is that all Socrates is saying by insisting so strongly on his own ignorance, even his ignorance of virtue?

Even if we grant Vlastos that Socrates did implicitly distinguish between provisional and definitive knowledge, it is hard to shake a suspicion that something important is lost in this interpretation. He acknowledges it himself at the end of his essay: does this really explain, Vlastos wonders, why Socrates would be so surprised by the oracle's pronouncement of his surpassing wisdom? Socrates was recognized by everyone as a master of dialectic; he himself never doubted his skill as an examiner. If the Pythia's point was just that he was as wise as human beings were capable of being, given the impossibility of certain knowledge, would that really have caused the sort of existential crisis that reshaped Socrates's life? Vlastos attributes the discrepancy to the fact that Socrates was not finally an epistemologist (this is a very elementary epistemological distinction, after all) but a "moralist pure and simple," awed with a religious awe at the distance between human wisdom and the wisdom of the gods.[12] Perhaps, in part. But what Vlastos ignores is precisely the *moral* significance of ignorance, its place in a well-lived life.

Ignorance is never a merely epistemological category for Socrates. It is always first and foremost a moral one. And it is an exceedingly complicated, even paradoxical moral category; it would take a much more exegetically technical essay than this one to unpack it properly. On the one hand, he indicts his opponents for their ignorance, and often treats ignorance as essentially synonymous with wrongdoing. "No one who has a particle of understanding" could agree with Meletus's empty and contrived accusations against him, Socrates says in the

Apology.[13] "Injustice is ignorance—no one could still not know this," he says in the *Republic*.[14] But on the other hand, the ignorance he attributes to himself is essential to his virtue. "Surely it is the most blameworthy ignorance to believe that one knows what one does not know," he says,[15] and so surely it is better—morally better, better for one's soul—to embrace one's own ignorance when ignorant is what one is. It's for just this reason that he commits himself to this quixotic quest of convincing everyone he meets that they know nothing.

Consider the start of the *Protagoras*—a dialogue, by the way, in which Socrates later argues, if inconclusively, that virtue is reducible to wisdom and wrongdoing to ignorance. A man named Hippocrates comes knocking excitedly on Socrates's door before the sun is up. He cannot wait to tell Socrates the news: Protagoras, the celebrity sophist, has come to Athens. Why are you so anxious to meet him?, Socrates asks. Has he hurt you in some way? Yes, Hippocrates says, only half joking: "he alone is wise, but he doesn't make me wise too!"[16] Socrates voices his skepticism—someone like Protagoras will be happy to make you wise, if only you empty your pockets for him—but he agrees to go along and help make the introduction.

As they walk, Socrates probes Hippocrates himself, getting him to think about what it is he wants out of a teacher like Protagoras. It's obvious what we want out of a reading teacher or a math teacher, but what does someone calling himself a "sophist" have to offer? It's right there in the name, surely: the *sophistēs* knows *tōn sophōn*, wise things. But this is vacuous. The painter and the sculptor know "wise things" too. The question is which wise things the sophist is wise about. About speaking cleverly, maybe?[17] Same problem: speaking cleverly *about what?* Hippocrates can't say (and can't bring himself to say that he doesn't know). "What then? Do you know what kind of danger you're putting your soul into?"[18] The question is thick with irony. Hippocrates doesn't even know what a sophist is or what this sophist teaches, so of course he can't know the danger he might be walking into.

I'm going to suggest, in a moment, that we stand in the same relationship to "artificial intelligence" that Hippocrates does to Protagoras—we cannot say what kind of intelligence we find in it, and so we face the same danger. But first let me make two more observations about the sort of ignorance that Socrates is commending.

First, notice *how* Socrates goes about his inquiry with both Hippocrates and Protagoras. This is well-known to anyone familiar with Socrates's style, but since our interest is in ethics pedagogy, it's worth describing concretely. Socrates opens the conversation with a point that his interlocutor *unknowingly claims to know*—which is to say, a claim to knowledge implicit in his interlocutor's action but not explicitly recognized. In his rush to become Protagoras's student, Hippocrates assumes he knows what Protagoras can teach him, but it becomes clear on examination that he does not. When Socrates comes to examine Protagoras himself, he focuses on the question of whether virtue is one thing or several, a

question that someone who claims to teach virtue should surely know. Again, it becomes clear that Protagoras does not. Characteristically, neither question (what the sophist knows, or whether virtue is one) gets answered. The answer wasn't the point. The point, rather, was to help the two men see that they do not know what they implicitly claim to know, which is just how Socrates describes his purpose in the *Apology*.[19] Hippocrates shows a hint of shame when Socrates reveals his ignorance to him (Socrates notices him blushing in the morning light); Protagoras, significantly, shows no shame at all but only stubbornness and anger. Hippocrates, after all, had already confessed that he was only seeking wisdom, which is why he was so eager to see Protagoras in the first place. But Protagoras had staked his entire social and professional identity on already being wise.

Second, notice *why* Socrates thinks it important to try to convince Hippocrates and Protagoras of their own ignorance. We have already seen why he thinks it so important in Hippocrates's case: because in his willingness to throw all his money at Protagoras and submit to his teaching, Hippocrates is wagering his own soul that Protagoras will make him a better person. Admitting his ignorance about whether Protagoras can do that—admitting his ignorance about what Protagoras even knows—means that, rather than rushing headlong, he can do the very sensible thing that Socrates goes on to do on his behalf: to *ask* Protagoras to explain what he has to offer, and then to test whether what he offers really is good. There is thus an immediate practical payoff to recognizing his ignorance. Knowing one's own ignorance is a way of caring for one's soul.

What about Protagoras himself? Socrates shows less interest in Protagoras's soul, but only because Protagoras has raised the ante. Pitching his services, Protagoras tells Hippocrates that "on the very day you join me, you will go home having become a better person, and the same on the next day."[20] Socrates asks him the same question he had asked Hippocrates: better in what way? Protagoras clarifies: he teaches *euboulia*, how to deliberate well, both about one's own affairs and about the affairs of the city. He teaches the art of *polis*-building; he fashions good citizens.[21] So Protagoras is wagering not only his own soul but the soul of the city itself on the conviction that he knows what good judgment is, what good citizenship is, what virtue itself is. Protagoras never backs down from the claim that he does know those things, even when confronted with his own inconsistencies, so we cannot know what he might have done if he accepted his ignorance. We are left to wonder and worry about what will happen when he forges recklessly ahead, teaching moral deliberation even though he is ignorant of it. His unknowing ignorance imperils the whole city.

AI and Moral Deliberation

Like Protagoras, university ethics teachers often claim to teach *euboulia*. We claim to teach students the art of moral deliberation, for their own sake and for the sake of the *polis*. Presumably that means we know something about what it

means to deliberate well. Do we? Are we sure, for that matter, that the art of deliberating well is what university ethics teachers ought to teach?

I do not know very much about artificial intelligence (just how little I know will soon become clear, I'm sure) but it seems to me that the excitement over ChatGPT is, at least from a computational point of view, well deserved. The history of artificial intelligence is full of hype and premature celebration, so I trust I can be forgiven my initial skepticism. But as Ari Schulman argues—himself a former AI researcher and now editor of the technology-critical journal *The New Atlantis*—this might be the breakthrough people have been waiting for.[22] It has its clear limits, of course, but the program is capable of accomplishing a stunning range of tasks with stunning degrees of competence. It is impressively sensitive to context. It recognizes and produces natural language in a way often indistinguishable from human beings. These are things that no earlier species of AI has been able to do, at least not over so broad a range of conditions.

It is especially the generality and the language sensitivity of large language models that makes them *seem* like a technology that might be useful in an ethics classroom. In particular, it seems to be able to do the thing that Protagoras claimed to teach. It seems to be able to *deliberate*. ChatGPT can provide plausible answers, of a sort, to moral questions; it can pick out, apparently, morally salient features of a situation. When I ask ChatGPT whether the U.S. should institute a universal basic income, for example, I'm told that it's a complex question that encompasses matters of politics, culture, and economics. I'm told that proponents think it could alleviate poverty and provide a safety net for people losing their jobs to (poignantly) artificial intelligence, while critics think that it could drive up inflation and disincentivize work. These are the kinds of considerations many of us would hope to see appear in student essays, and expressed more clearly than many of us would hope to see students express them.

What ChatGPT does not tell me, notably, is whether the U.S. should institute a universal basic income. The program categorically refuses to issue a judgment on the question, or on any moral question I can think to ask it. Even bland and uncontroversial questions, like whether it is good to be nice to others, are answered with subtle redirections about what "most moral frameworks" would hold. Nor does the program propose any suggestions about how to sort through the various moral considerations that it has identified.

Although I do not know how programmers have fine-tuned ChatGPT when it comes to dealing with moral questions, its refusal to issue judgments is very likely a guardrail explicitly intended by the programmers. That the programmers thought it was important to put such guardrails in place is interesting in its own right. But an *incapacity* to issue such judgments, at least in the sense that we usually mean when we talk about moral decision-making, is part of the program design itself. All large language models, after all, are essentially "autoregressive token predictors":[23] they work not by gathering evidence or considering the coherence of claims or assigning priority to different considerations—in short, not

by doing any of the things we usually think of as part of practical judgment—but by making educated guesses about what piece of a word is most likely to follow in the sentence it is writing.[24] Whether or not it is true that *all* artificial intelligence is reducible to applied statistics, as Ted Chiang has suggested,[25] large language models certainly are. What looks like subtle contextual awareness on the computer's part is, at root, an incomprehensibly (to a human) long list of vectors representing the likelihood of a token's occurrence in combination with other tokens in its training data. What looks like facility with natural language is, at root, an incomprehensibly (to a human) long list of mathematical operations between those vectors. "Programs like ChatGPT," Cal Newport says, "don't represent an alien intelligence with which we must now learn to coexist; instead, they turn out to be run on the well-worn digital logic of pattern-matching, pushed to a radically larger scale."[26]

Our inability to follow or explain the many billions of operations that large language models are performing on these vectors is what gives them their "black box" quality. We cannot explain, concretely, how it is that ChatGPT is producing its answer to my question about a universal basic income—and thus far, ChatGPT can't explain how it's producing its answer either. This interpretability problem has emerged as a key question in the ethics of AI.[27] Is it responsible to rely on black-box systems for decisions about what treatment to pursue when we do not know on what grounds the system is recommending that treatment, for example?[28] How about for decisions about criminal sentencing or parole, where algorithms have been shown to reflect racial bias?[29]

There are hard questions to consider about what sorts of interpretability we can or should demand of these programs,[30] as well as what sorts of interpretability are even technically possible. But in relation to moral deliberation—which, again, is the main thing we and Protagoras say we're trying to teach—the problem is not that AI's moral judgments are uninterpretable but that *AI is not making moral judgments at all*. Socrates could not examine ChatGPT. ChatGPT is a token predictor, not a practical deliberator. The reasons it could give for what it says, if it could be made to give reasons, would have exclusively to do with statistical regularities in the text of its training data. Whatever our particular normative account of practical judgment, however we hope that our students (and we ourselves) go about making practical judgments as a result of our classes, token prediction plays no part in it. At the level of the whole system, a sort of transparency that Zachary Lipton calls "simulatability,"[31] ChatGPT and its ilk are actually not especially opaque. We can understand and even mimic what large language models are doing, if with far, far less statistical sophistication. The problem with using ChatGPT in the context of moral deliberation is not that it is a black box; the problem is that ChatGPT is doing something fundamentally unlike moral deliberation.

So when ChatGPT seems to be picking out morally salient issues having to do with instituting a universal basic income, what it is really doing is telling

me that words like "unemployment" and "incentive" and "inflation" frequently occur in close connection with the words "universal basic income" in its training data. Or more generously, from a wider point of view, we might say that it is reflecting back to us that most texts in its training data that discuss a universal basic income speak for or against it in these terms. It is not making a judgment about what issues are in fact morally salient. But even if we grant that ChatGPT is not making moral judgments of its own, might this more limited task still be useful in our work of making moral judgments?

Maybe. We might think of ChatGPT's answer to a moral question as a report on common opinion. Knowing what others have thought about a question is certainly not a bad thing. On the contrary, one of the first things that I myself do when considering a new question is to search out a guide to the conversation, someone who can sum up the way others have parsed the issue. I encourage my students to do the same. Maybe this is what ChatGPT can offer.

The specter of common opinion always haunts the ethics classroom. It haunts every moral judgment. While Socrates was sitting in prison, awaiting execution, his friend Crito came early in the morning to try to persuade him to escape. One of the reasons Crito gave—and he gave many reasons, all jumbled up with another, not knowing what would stick—was that his reputation would be destroyed if people thought that he was too cheap to pay off the guards. "But my happy Crito," Socrates answers, "why do we care so much about the opinion of the many?"[32] Crito points out in reply that what the many think matters very much; Socrates's imprisonment is proof enough of that. But Socrates demurs. The many are capable neither of great evils nor of great goods. They are too fickle for either. "They cannot make a person either thoughtful or thoughtless. They just do whatever they happen to do."[33] Against both the hype and the apocalyptic warnings about ChatGPT, I suspect we should say the same about this current breed of artificial intelligence. It captures some true associations; it also makes very basic mistakes. It is thus no kind of guide to a good life, and it is often an unreliable guide even to that phantasmic thing we call common opinion. As Socrates tells Crito and as ethics teachers tell their students, what finally matters is not what common opinion holds but what we are persuaded is in fact good and true.

Even as a preliminary summary of common opinion, to be used, tested, honed and surpassed—in a mode more Aristotelian than Socratic, perhaps—there is reason to be wary of ChatGPT's usefulness. What these large language models offer, to invoke Ted Chiang again, is a "blurry JPEG of the web."[34] What they give us is the result of a lossy compression algorithm, a loose paraphrase of a million forgotten originals. When we ask our students to provide summaries of a conversation, note, we often ask them to be able to tell us where they learned particular pieces of information. ChatGPT cannot do that, since every particular source has been reduced to an anonymous datapoint. There is a veneer of objectivity in this. Every bit of input receives equal weight. But presumably we

ask our students to cite their sources because we want them to learn to exercise judgment about which sources are more and less trustworthy, a sort of judgment that ChatGPT does not and cannot exercise. Is it then the right sort of guide even to the *endoxa*? Even a summary of received opinion can be better or worse, more or less helpful. The very best summaries themselves bear the mark of a piercing intelligence, organizing the conversation in a way that is pedagogically helpful or that brings central conceptual issues to the fore. Many entries in the Stanford Encyclopedia of Philosophy, for example, are excellent in just this way. When I read an entry on an issue that is new to me, I come away not only with a sense of the major sites of debate but also, more importantly, with an illuminating map of the territory. It is far from clear that a blurry JPEG is the sort of map we should want to offer our students.

Exercises in Ignorance

I have so far avoided saying what moral deliberation is, trying only to point out that what ChatGPT is doing is not deliberation or even an especially helpful preliminary to it. I still won't define it, because I do not know how. And anyway, Socrates gives us reason to wonder whether teaching moral deliberation is quite so central to the task of the ethics teacher as we usually suppose.

In the *Symposium*—and I am admittedly venturing beyond the properly Socratic dialogues here, but I think Plato is capturing something important about Socrates's intellectual posture—Socrates chides his fellow dinner-party guests for exaggerating the virtues of love. Even in praising something, we ought to tell the truth about it. Love is defined, Socrates counterintuitively suggests, by the *lack* of what it desires. Who yearns for something she already has? The philosopher, the lover of wisdom, is therefore not the one who has wisdom, but the one who longs for it, and longing for it, knows that she does not have it. The philosopher, by definition, knows that she is ignorant.

Love is not the most beautiful and best, as his friend and host Agathon had said it was; it is a *desire* for what is most beautiful and best. But that does not mean, Socrates remembers Diotima persuading him, that it must therefore be something ugly and bad. "Don't force what is not beautiful to be ugly," she says, "or what is not good to be bad. For it is this way with love, too. When you agree that he is neither good nor beautiful, do not think he must be ugly or bad instead, but something in between."[35] Love is strange, perhaps somewhat tragic. Love is caught between what it is and what it reaches out for. "This is the difficult thing about ignorance," Diotima says later: "though you are neither beautiful nor good nor intelligent, you seem to yourself to be enough. If you think you need nothing, you won't want what you don't think you need."[36] It certainly not this sort of ignorance that we want for our students. What we want for them is the philosophical sort, the ignorance that is a form of wisdom because it desires the wisdom it knows it does not possess.

It is not easy to know how to teach this sort of wisdom, though if Plato is to be believed, Socrates did somehow manage to inspire it in his students. They became enamored of the very pursuit of knowledge, even if—perhaps even because—they knew their desire for it would never be fulfilled.[37] It is not easy to measure this sort of learning, either, except over the course of a long life. I suspect it is these practical difficulties that keep us university ethics teachers from taking Socrates's style too seriously. We want discrete and assessable learning outcomes (and even if we don't want them, we work in an educational bureaucracy that requires them). A proceduralist notion of moral deliberation satisfies that desire much more neatly. We can distinguish key steps that we take to define deliberation—reflection on one's own presuppositions, a fair recital of other points of view, the application of a particular theory to a particular problem, and so on—and grade our students on how explicitly and thoroughly they have performed those steps. The possibility that competence in such procedures has little relationship to being a good person, and may actually better correlate with being a good *rationalizer* or *manipulator*, is inconvenient and usually ignored.[38]

Proceduralist rationality is also where computers excel. In fact, it is the only sort of intelligence that computers really possess. Although AI built on large language models themselves is not deliberating about goods even in a proceduralist way, one could imagine some other form of AI at some other point in the future doing so. What is much more difficult to imagine is a computer who is a philosopher, knowing it knows nothing and longing for knowledge.

One of the most jarring things about working with ChatGPT is its indifference to the truth. It invents dates and numbers and sources and quotes, all without giving any hint that it is inventing. And this is because, by design, the program is not aiming at giving correct answers to a question; it is aiming at making statistically probable predictions about a sequence of tokens. The most generous interpretation would be that the truth of its claims is "emergent" from its linguistic sophistication, a function of an unthinkably complex model of how the words in its training data are interrelated. The more straightforward interpretation is that it only tells the truth on accident. The program does not know or care to know whether what it is saying is true. ChatGPT is therefore, in Harry Frankfurt's technical sense, a bullshit machine.[39] What distinguishes a bullshitter from a liar, Frankfurt says, is what exactly they misrepresent. The liar is trying to misrepresent reality, trying to deceive someone about what he takes to be the truth of the matter. The bullshitter is unconcerned with the truth of the matter; he is trying to misrepresent *himself*. The bullshitter—like ChatGPT's fine-tuned public-facing persona—presents himself as someone who knows and cares to know the truth, but in fact does not. "The bullshitter is faking things," Frankfurt says. "But this does not mean that he necessarily gets them wrong."[40] Whether or not he tells the truth is beside the point.

Or maybe, to bring the comparison back to ancient Athens, we could think of ChatGPT as a kind of sophist. Socrates's enemies accused him of "making the

weaker argument stronger,"[41] playing with words in order to convince others of their intelligence and manipulate others to their advantage. In Aristophanes's *Clouds*, a poor man named Strepsiades, desperate to rid himself of crushing debt, goes to Socrates's "think-factory of wise souls"[42] in hopes that he might learn how to bamboozle his debtors into forgiving him. Strepsiades says that "these men teach—if you give them money—how to defeat others by speaking, both justly and unjustly."[43] ChatGPT will also give you arguments for any position you ask it to defend—barring those that its programmers have decided are beyond the pale—and if you are willing to pay a low monthly fee, it will give you even better ones.

I do not know whether future iterations of this technology will be better able to measure its output against other established databases of knowledge. I hope it is, and many smart people are already hard at work at making it so. But even if it does, it will be exhibiting a qualitatively different kind of intelligence than Socrates embodied—a qualitatively different kind of intelligence than I want for my students. As the famed classicist Pierre Hadot puts it, "in Socratic dialogue, the real question in play is not *that of which one speaks* but *the one who is speaking*."[44] Socrates's teaching, to the extent that we can call it that, is aimed at turning his interlocutors' attention back onto themselves. We saw this clearly in the *Protagoras*, in Socrates's plaintive question to Hippocrates: don't you know what danger you're putting your soul into? Your presumption of knowledge dooms you. Socrates is not trying to teach any concrete thing at all, not even the practice of deliberation. Certainly "Socrates has no system to teach," Hadot says in a different essay. "His philosophy is entirely a spiritual exercise, a new mode of life, active reflection, living conscience."[45]

Socrates, to repeat, wanted to convince people of their ignorance. This was not a merely deflationary exercise, born of cynical resentment. He was not merely trying to bring people down a peg. He thought that an honest recognition of one's own ignorance was fundamental to human wisdom, fundamental to the care of oneself and others, fundamental to a form of life oriented by the love of truth and justice. I suppose it goes without saying that ChatGPT does not know its own ignorance, and can't be convinced of it. That is because, as I have already suggested, it is programmatically disinterested in the truth-value of its statements. What might look like an avowal of ignorance—its frequent reminders that it knows nothing more recent than the completion of its training, for example, or that it has no opinions about what public policies are better or worse—in fact distracts us from its deeper ignorance. ChatGPT is in some ways more like the poets or the tradesfolk in the *Apology* who, because they know one thing well, convince themselves of their skill in other areas: ChatGPT is undeniably good at token prediction, but token prediction is not the same as knowledge or, certainly, wisdom. ChatGPT, lacking spirit (*psychos*), is obviously not a good candidate for Hadot's spiritual exercises. And it should be equally clear that ChatGPT, not knowing its own ignorance, is incapable of teaching others to

engage in such spiritual exercises. If this sort of AI is a bad guide to moral deliberation, it is an even worse guide to this deeper sort of moral inquiry.

The language of "spiritual exercises" is uncomfortable for many contemporary readers, and even more uncomfortable for contemporary ethics teachers. The idea strikes many as too mystical, too religious, too ethereal to be of much use in the university classroom.[46] Hadot is aware that the term is confusing, but insists that no substitute will do if we want to understand the full scope of what Socrates and his Hellenistic heirs were trying to do.[47] To focus just on the "intellectual," or just on the "moral," or just on the "psychological" would truncate things. The goal of Socratic philosophy is nothing less than the transformation of the whole person. Such a goal is impossibly ambitious, but that is part of the point. This sort of pedagogy focuses on giving students something to strive for, even if that thing is finally unreachable, by reframing one's sense of self in relation to the whole.[48] The practice of ignorance is primary because I cannot set out to a final destination if I think I have already arrived there.

Nor is this sort of pedagogy merely hortatory. It does not consist in sermonizing to students about the dangers of overconfidence. As Hadot makes clear, the spiritual exercises that the Hellenistic philosophers recommended were quite concrete: practices of self-awareness and self-analysis, confrontation with and meditation on death, concentration on the present moment, memorization of key axioms. Most important of all, especially in relation to the practice of ignorance, is the exercise of *dialogue*, examining the presuppositions of others and submitting your own presuppositions to examination. Even private meditation, Hadot emphasizes, is essentially dialogical, as Socrates demonstrates when he allows himself to be examined by the laws of Athens in the *Crito* or by Diotima in the *Symposium*. Socrates does not adopt this method because he thinks that dialogue is good practice for participation in a democratic citizenry, or because he thinks that open consideration of multiple points of view is the best way to truth.[49] He adopts it because, rightly practiced, it exposes a person to his own ignorance.

We can be more concrete. One of Socrates's most consistent strategies is to focus the discussion on some matter of immediate, existential concern. He talks to Euthyphro about the meaning of holiness because it is on that basis that Euthyphro plans to prosecute his father. He talks to Hippocrates about teachers because Hippocrates is about to hire one. He talks to Gorgias about rhetoric because Gorgias advertises himself as a rhetorician. Socrates's goal is not to convince them that they are ignorant about everything—though that is what he claims for himself—but that they are ignorant of the specific things that their action commits them to being knowledgeable about. It is this sort of ignorance in particular that Socrates takes as a crucial part of caring for themselves and others. In one of his more direct moments, Socrates accuses his accuser, Meletus, saying, "He says I do wrong by corrupting the young. But I, fellow Athenians, say that Meletus does wrong, because he jokes around with earnestness, readily

involving people in lawsuits, pretending to be earnestly concerned and troubled about things he never cared about at all."[50] Meletus's ignorance of his own ignorance has led him to injure Socrates, of course, but more importantly, it has led him to injure himself and all of Athens, by driving them to put a man unjustly to death.[51] The sort of inquiry that Socrates is modeling here, identifying and exposing ignorance about matters of transparent practical concern, is something ethics teachers can certainly continue to do.

The point of this dialogical self-examination, as Hadot has emphasized, is not ultimately to help students develop a stronger, more defensible basis from which to act. The dialogue is formative in itself, because it shapes people into truth-lovers, people who strive for the wisdom they know they lack—or else exposes them as people who care more about money or prestige or something else than they do about truth. It creates philosophers, but also distinguishes philosophers from bullshitters.

This is perhaps an anticlimactic conclusion, after everything I have said so far, but here it is. An artificial intelligence like that embodied in ChatGPT and similar technologies simply does not seem to have anything special to contribute to the work of teaching or practicing ethics, whether conceived in the Protagorean sense as *euboulia*, sound deliberation, or as Socratic ignorance. Better to read books and examine them; better to have conversations and examine one another; better to meditate or journal and examine oneself. It is no affront to the computational genius that went into creating these programs to say that they are unimportant to the work of self-examination. For this work one needs only another spirit willing to inquire after the truth, at least for a while.

Notes

1. An earlier version of this paper was presented at the 24th Annual Conference on Ethics Across the Curriculum, James Madison University, Harrisonburg, Virginia, October 1–3, 2023. I am grateful for the clarifying comments and questions I received there.

2. This story is told in Plato's *Apology*, beginning at 20e. All quotations from Plato's dialogues in this essay are my own translations, based on the Greek of the Oxford Classical Texts. The standard collection of English translations is Plato, *Complete Works*, ed. John M. Cooper and D.S. Hutchinson (Indianapolis, IN: Hackett, 1997).

3. In fact, he does not present even this as settled doctrine: his commitment to endlessly questioning anyone with a claim to wisdom is framed precisely as an attempt to *disprove* the oracle's claim that Socrates himself is the wisest. He seems to

hold open the possibility that he may be ignorant even about the value of his own ignorance.

4. The contrast I draw between the Socratic and Protagorean modes of teaching ethics could easily be complicated, of course. It is not hard to imagine Protagoras accepting a certain (less radical) version of Socrates's embrace of ignorance, nor is it hard to imagine Socrates encouraging his hearers to practice sound deliberation. In fact, Socrates could be read as offering something like his own version of *euboulia* towards the end of the *Protagoras*, when he presents wisdom as an "art of measurement" (*Prot*. 356e). But I leave them distinct here both because Socrates himself thought they were distinct and because it is heuristically helpful to deal with a clear option.

5. Apol. 23a: οἴονται γάρ με ἑκάστοτε οἱ παρόντες ταῦτα αὐτὸν εἶναι σοφὸν ἃ ἂν ἄλλον ἐξελέγξω.

6. Plato paints a vivid picture of this tactic in the *Euthydemus*.

7. Apol. 31b–31c.

8. Alexander Nehamas, *Virtues of Authenticity: Essays on Plato and Socrates* (Princeton, NJ: Princeton University Press, 1999), ch. 2.

9. The question of whether virtue can be taught is a recurring theme in the Socratic dialogues, appearing most clearly in the *Meno* and the *Protagoras*.

10. Gregory Vlastos, "Socrates' Disavowal of Knowledge," *The Philosophical Quarterly* 35, no. 138 (1985): 1–31.

11. Apol. 20d.

12. Vlastos, "Socrates' Disavowal of Knowledge," 28.

13. Apol. 27e: ἄν καὶ σμικρὸν νοῦν ἔχοντα.

14. Rep. 351a: ἐστὶν ἀμαθία ἡ ἀδικία—οὐδεὶς ἂν ἔτι τοῦτο ἀγνοήσειεν.

15. Apol. 29b: ἀμαθία ἐστὶν αὕτη ἡ ἐπονείδιστος, ἡ τοῦ οἴεσθαι εἰδέναι ἃ οὐκ οἶδεν.

16. Prot. 310d: μόνος ἐστὶ σοφός, ἐμὲ δὲ οὐ ποιεῖ.

17. The Greek word used for "clever" here, *deinos*, also means "terrible," which seems fitting.

18. Prot. 313a: τί οὖν; οἶσθα εἰς οἷόν τινα κίνδυνον ἔρχῃ ὑποθήσων τὴν ψυχήν;

19. Apol. 23b.

20. Prot. 318a: ᾗ ἂν ἡμέρᾳ ἐμοὶ συγγένῃ, ἀπιέναι οἴκαδε βελτίονι γεγονότι, καὶ ἐν τῇ ὑστεραίᾳ ταὐτὰ ταῦτα.

21. Prot. 318e–319a.

22. Ari Schulman, "Why This AI Moment May Be the Real Deal," *The New Atlantis* (https://www.thenewatlantis.com/publications/why-this-ai-moment-may-be-the-real-deal, 2023).

23. I borrow this shorthand from Cal Newport, "The End of Screens?" *CalNewport.com*, May 2023.

24. In trying to understand how large language models work, I have benefitted greatly from Stephen Wolfram, "What Is ChatGPT Doing . . . and Why Does It Work?" *Stephen Wolfram: Writings* (https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/, February 2023), and Timothy B. Lee and Sean Trott, "A Jargon-Free Explanation of How AI Large Language Models Work," *Ars Technica* (https://arstechnica.com/science/2023/07/a-jargon-free-explanation-of-how-ai-large-language-models-work/, July 2023).

25. Madhumita Murgia, "Sci-Fi Writer Ted Chiang: 'The Machines We Have Now Are Not Conscious'," *Financial Times*, June 2023.

26. Cal Newport, "What Kind of Mind Does ChatGPT Have?" *The New Yorker*, April 2023.

27. For a compact and helpful summary of this conversation, if now a few years out of date, see S. Matthew Liao, ed., *Ethics of Artificial Intelligence* (New York: Oxford University Press, 2020), 7–9.

28. Alex John London, "Artificial Intelligence and Black-Box Medical Decisions: Accuracy Versus Explainability," *Hastings Center Report* 49, no. 1 (2019): 15–21.

29. Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York, NY: St. Martin's Press, 2018); Julia Dressel and Hany Farid, "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances* 4, no. 1 (January 2018).

30. Zachary C. Lipton, "The Mythos of Model Interpretability," *Communications of the ACM* 61, no. 10 (September 2018): 36–43.

31. Lipton, 40.

32. Crit. 44c: ἀλλὰ τί ἡμῖν, ὦ μακάριε Κρίτων, οὕτω τῆς τῶν πολλῶν δόξης μέλει;.

33. Crit. 44d: οὔτε γὰρ φρόνιμον οὔτε ἄφρονα δυνατοὶ ποιῆσαι, ποιοῦσι δὲ τοῦτο ὅτι ἂν τύχωσι.

34. Ted Chiang, "ChatGPT Is a Blurry JPEG of the Web," *The New Yorker*, February 2023.

35. Symp. 202b: μὴ τοίνυν ἀνάγκαζε ὃ μὴ καλόν ἐστιν αἰσχρὸν εἶναι, μηδὲ ὃ μὴ ἀγαθόν, κακόν. οὕτω δὲ καὶ τὸν ἔρωτα ἐπειδὴ αὐτὸς ὁμολογεῖς μὴ εἶναι ἀγαθὸν μηδὲ καλόν, μηδέν τι μᾶλλον οἴου δεῖν αὐτὸν αἰσχρὸν καὶ κακὸν εἶναι, ἀλλά τι μεταξύ.

36. Symp. 204a: αὐτὸ γὰρ τοῦτό ἐστι χαλεπὸν ἀμαθία, τὸ μὴ ὄντα καλὸν κἀγαθὸν μηδὲ φρόνιμον δοκεῖν αὑτῷ εἶναι ἱκανόν. οὔκουν ἐπιθυμεῖ ὁ μὴ οἰόμενος ἐνδεὴς εἶναι οὗ ἂν μὴ οἴηται ἐπιδεῖσθαι.

37. Plato's description of Alcibiades's mad love for Socrates suggests that maybe he thought the inaccessibility of what one desires, including knowledge, inflames one's desire for it all the more.

38. Jonathan Haidt, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Review* 108, no. 4 (October 2001): 814–34, has become a locus classicus for the argument that rational

deliberation is, empirically speaking, a post-hoc defense of moral judgments we have already made on other grounds. For a later and longer version of the argument, see Jonathan Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion*, Reprint edition (New York: Vintage, 2013).

39. Harry G. Frankfurt, "On Bullshit," in *The Importance of What We Care About: Philosophical Essays* (New York: Cambridge University Press, 1988), 117–33. The aptness of Frankfurt's category of bullshit struck many observers very quickly. The first reference I've been able to find is Arvind Narayanan and Sayash Kapoor, "ChatGPT Is a Bullshit Generator. But It Can Still Be Amazingly Useful," *AI Snake Oil* (https://www.aisnakeoil.com/p/chatgpt-is-a-bullshit-generator-but, December 2022), published just a week after the public release of ChatGPT. It has started to appear in academic analysis of ChatGPT, too, in Eamon Costello, "ChatGPT and the Educational AI Chatter: Full of Bullshit or Trying to Tell Us Something?" *Postdigital Science and Education*, March 2023, and Nandita Roy and Moutusy Maity, "'An Infinite Deal of Nothing': Critical Ruminations on ChatGPT and the Politics of Language," *DECISION* 50, no. 1 (March 2023): 11–17.

40. Frankfurt, "On Bullshit," 129.

41. Apol. 19b: τὸν ἥττω λόγον κρείττω ποιῶν

42. S. Douglas Olson, *Aristophanes' Clouds: A Commentary*, (Ann Arbor: University of Michigan Press, 2021), line 94: ψυχῶν σοφῶν . . . φροντιστήριον. My translation.

43. Olson, *Aristophanes' Clouds*, lines 98–99: οὗτοι διδάσκουσ', ἀργύριον ἤν τις διδῷ, / λέγοντα νικᾶν καὶ δίκαια κἄδικα.

44. Pierre Hadot, *Exercices spirituels et philosophie antique, nouvelle édition revue et augmentée*, 41 (Paris: Michel, 2002), 39, my translation: Dans le dialogue socratique, la vraie question qui est en jeu n'est pas ce dont on parle, mais celui qui parle.

45. Hadot, 118: Socrate n'a pas de système à enseigner. Sa philosophie est tout entière exercice spirituel, nouveau mode de vie, réflexion active, conscience vivante.

46. The idea has even proven uncomfortable for supportive readers of Hadot. Thus the title of Hadot's most famous book, *Exercices sprituels et philosophie antique*, becomes in English, *Philosophy as a Way of Life*; and the title essay, "Exercices spirituels," first in the French edition, gets demoted to third position in the English edition. See Pierre Hadot, *Philosophy as a Way of Life* (Malden, MA: Wiley-Blackwell, 1995).

47. Hadot, *Exercices spirituels et philosophie antique*, 20–21.

48. Julia Annas, "The Sage in Ancient Philosophy," in *Anthropine Sophia*, ed. Francesca Alesse and Gabriele Giannantoni (Napoli: Bibliopolis, 2008), 11–27 offers a complementary account of the role of the ideal of the sage in Hellenistic philosophy.

49. Stephen Brookfield and Stephen Preskill, *Discussion as a Way of Teaching: Tools and Techniques for Democratic Classrooms*, 2nd ed (San Francisco: Jossey-Bass, 2005), for example, adopt something like this rationale. "Discussion is one of the best ways to nurture growth because it is premised on the idea that only through

collaboration and cooperation with others can we be exposed to new points of view . . . . In the process, our democratic instincts are confirmed: by giving the floor to as many different participants as possible, a collective wisdom emerges that would have been impossible for any of the participants to achieve on their own" (4).

50. Apol. 24c: φησὶ γὰρ δὴ τοὺς νέους ἀδικεῖν με διαφθείροντα. ἐγὼ δέ γε, ὦ ἄνδρες Ἀθηναῖοι, ἀδικεῖν φημι Μέλητον, ὅτι σπουδῇ χαριεντίζεται, ῥᾳδίως εἰς ἀγῶνα καθιστὰς ἀνθρώπους, περὶ πραγμάτων προσποιούμενος σπουδάζειν καὶ κήδεσθαι ὧν οὐδὲν τούτῳ πώποτε ἐμέλησεν.

51. Apol. 30c–d.

References

Annas, Julia. "The Sage in Ancient Philosophy." In *Anthropine Sophia*, edited by Francesca Alesse and Gabriele Giannantoni, 11–27. Napoli: Bibliopolis, 2008.

Brookfield, Stephen, and Stephen Preskill. *Discussion as a Way of Teaching: Tools and Techniques for Democratic Classrooms*. 2nd ed. San Francisco: Jossey-Bass, 2005.

Chiang, Ted. "ChatGPT Is a Blurry JPEG of the Web." *The New Yorker*, February 2023.

Costello, Eamon. "ChatGPT and the Educational AI Chatter: Full of Bullshit or Trying to Tell Us Something?" *Postdigital Science and Education*, March 2023.

Dressel, Julia, and Hany Farid. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances* 4, no. 1 (January 2018).

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press, 2018.

Frankfurt, Harry G. "On Bullshit." In *The Importance of What We Care About: Philosophical Essays*, 117–33. New York: Cambridge University Press, 1988.

Hadot, Pierre. *Exercices spirituels et philosophie antique*. Nouvelle édition revue et augmentée. 41. Paris: Michel, 2002.

Hadot, Pierre. *Philosophy as a Way of Life*. Malden, MA: Wiley-Blackwell, 1995.

Haidt, Jonathan. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108, no. 4 (October 2001): 814–34.

Haidt, Jonathan. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Reprint edition. New York: Vintage, 2013.

Lee, Timothy B., and Sean Trott. "A Jargon-Free Explanation of How AI Large Language Models Work." *Ars Technica*. https://arstechnica.com/science/2023/07/a-jargon-free-explanation-of-how-ai-large-language-models-work/, July 2023.

Liao, S. Matthew, ed. *Ethics of Artificial Intelligence*. New York: Oxford University Press, 2020.

Lipton, Zachary C. "The Mythos of Model Interpretability." *Communications of the ACM* 61, no. 10 (September 2018): 36–43.

London, Alex John. "Artificial Intelligence and Black-Box Medical Decisions: Accuracy Versus Explainability." *Hastings Center Report* 49, no. 1 (2019): 15–21.

Murgia, Madhumita. "Sci-Fi Writer Ted Chiang: 'The Machines We Have Now Are Not Conscious'." *Financial Times*, June 2023.

Narayanan, Arvind, and Sayash Kapoor. "ChatGPT Is a Bullshit Generator. But It Can Still Be Amazingly Useful." *AI Snake Oil*. https://www.aisnakeoil.com/p/chatgpt-is-a-bullshit-generator-but, December 2022.

Nehamas, Alexander. *Virtues of Authenticity: Essays on Plato and Socrates*. Princeton, NJ: Princeton University Press, 1999. Newport, Cal. "The End of Screens?" *CalNewport.com*, May 2023.

Nehamas, Alexander. "What Kind of Mind Does ChatGPT Have?" *The New Yorker*, April 2023.

Roy, Nandita, and Moutusy Maity. "'An Infinite Deal of Nothing': Critical Ruminations on ChatGPT and the Politics of Language." *DECISION* 50, no. 1 (March 2023): 11–17.

Schulman, Ari. "Why This AI Moment May Be the Real Deal." *The New Atlantis*. https://www.thenewatlantis.com/publications/why-this-ai-moment-may-be-the-real-deal, 2023.

Vlastos, Gregory. "Socrates' Disavowal of Knowledge." *The Philosophical Quarterly* 35, no. 138 (1985): 1–31.

Wolfram, Stephen. "What Is ChatGPT Doing . . . and Why Does It Work?" *Stephen Wolfram: Writings*. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/, February 2023.