



# From Artificial Intelligence to Artificial Wisdom: What Socrates Teaches Us

**Tae Wan Kim**, Carnegie Mellon University

**Santiago Mejia**, Fordham University

*A critical examination of existential questions may lead developers to design machines with higher forms of artificial intelligence. Infused by their ability to recognize their own ignorance, these machines would display not merely intelligence but wisdom.*

**E**ngineers, especially engineering students, should have an opportunity to think deeply about the nature of human flourishing and human excellence if they want to be educated to develop *good* artificial intelligence (AI). The conventional approach seeks to design AI that avoids causing

harm. But this approach falls short to the extent that it does not engage with the question “What is an excellent, flourishing, human being?” Socrates taught us two important things about this question: 1) reflecting on it was a central part of being human and 2) seriously engaging with this question leads to the recognition of a particular form of ignorance that is also a form of wisdom.

In this article, we will elaborate on these Socratic insights and show how they bear on AI. We hope that current and future engineers will be moved to build upon the ancient wisdom discussed here to reimagine their work on AI.

## VALUE ALIGNMENT AND HUMAN FLOURISHING

Businesses increasingly use AI to make important decisions for humans. Amazon, Google, and Facebook choose what users see. The driver-assist technology used in most brand-new vehicles aids drivers with steering and braking. Uber and Lyft match passengers with drivers and set prices. Though each of these examples comprises its own complicated technology, they share a core: a data-trained set of decision rules (often called machine learning or AI) that implements a decision with little or no human intermediation.

## EDITORS

**HAL BERGHEL** University of Nevada, Las Vegas; hlb@computer.org

**ROBERT N. CHARETTE** ITABHI Corp.; rncharette@ieee.org

**JOHN L. KING** University of Michigan; jlking@umich.edu



AI techniques are quickly being adopted to automate decisions. As this happens, societal worries about the compatibility of AI and human values grow. How can we ensure that AI does not turn against us? That it is under our control? That it serves us and promotes what we value? In response to these worries, some computer scientists have suggested that “value alignment” should be one of the top priorities in AI research.<sup>1,2</sup> Value alignment seeks to ensure that the technology we design incorporates the values that are important to us. The concept dates back to Alan Turing, who wrote about the need for machines to adapt to human values: “[T]he machine must be allowed to have contact with human beings in order that it may adapt itself to their standards.”<sup>3</sup>

The idea of value alignment is consistent with the IEEE’s approach to ethics. Recently, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems released an ambitious document outlining directives for “ethically aligned design.”<sup>4</sup> This document goes beyond the conventional approach that places the no-harm principle as a side constraint on design, emphasizing, instead, that human well-being and human flourishing should be central aims that technology should promote. Some technology developers have lost sight of this too often in recent years, often because their organizations have been too focused on the pursuit of short-term profits and bigger market share. Some of the most important problems to which technology has given rise, and which have turned public opinion distrustful of technological innovation, might have been averted by giving a more prominent role to human well-being and flourishing in the development of such innovations.

Putting human flourishing at the center of value alignment, however, is not simple. Offering a concrete and well-developed account of the nature of

flourishing can be seen as a fool’s errand. It is always challenging to offer such an account, but it is especially difficult to do so in a world with rapidly evolving technologies. Technologies are designed to solve a variety of human problems. In doing so, however, they inevitably reshape the ways in which humans interact with the world and flourish in it. The invention of the bow and arrow, for instance, enabled humans to hunt from a safe distance. This enabled them to reduce the risks of hunting at close range and expanded the availability of wild game. But the bow and arrow also modified the nature of hunting, thereby redefining what it meant to flourish as a hunter (and, given its application to war, to excel as a warrior). In sum, the difficulty in defining “flourishing” is not merely that it requires clarity about a host of central human concepts that are difficult to pin down but that it is in flux as technology opens and forecloses different kinds of existential possibilities.<sup>5</sup>

Is there any value to reflecting on human flourishing, given these difficulties? Socrates, the ancient figure, helps us to see why the answer is a resounding yes. He teaches us that recognizing that we fall short in articulating the nature of flourishing is a fundamental form of human wisdom. We propose that this form of Socratic wisdom should play a more prominent role in the development of AI.

## SOCRATIC IGNORANCE

During the trial at which he was condemned to death, Socrates explained how he had come to be “Athens’s gadfly.” An impulsive friend of his, Chaerophon, had asked the oracle of Delphi whether there was anyone wiser than Socrates. The oracle replied that no one was wiser. This response puzzled Socrates because he did not consider himself wise; he was aware that he did not have a well-worked-out account of the nature of human flourishing.

In an attempt to clarify the oracle’s meaning, Socrates sought those who were reputedly wise and asked them about their wisdom. He talked with politicians, poets, and craftsmen. After examining them through questions aimed to clarify their views and single out their implications, Socrates always reached the same conclusion: “[N]either of us knows anything worthwhile, but he thinks he knows something when he does not, whereas when I do not know, neither do I think I know. I am likely to be wiser to this small extent, that I do not think I know what I do not know.”<sup>6</sup> What Socrates thought was a form of ignorance turned out to be a form of wisdom.

Socratic wisdom, that is to say, Socratic ignorance, brings an increased openness and humility with respect to how the most important human questions should be answered. Socrates wanted to become wiser and did so by conversing with anyone about human flourishing, regardless of age, class, social status, and geographical origin. He did not exclude any view, no matter how apparently outrageous. Instead, he rigorously examined it in the hopes of learning from it. The fact that he was willing to examine everyone and that he was open to all sorts of opinions makes his approach a powerful tonic against echo chambers and filter bubbles. In addition, cultivating Socratic ignorance seems particularly important in a society like ours, where globalization is causing diverse cultures to clash and where technology is redefining, at a very fast pace, what it means to flourish as a human.

## SOCRATIC ENGINEERS

AI is a systematic approach to replicating human intelligence by using various mathematical, computational, and mechanical principles. The Turing test originated from “the imitation game,” in which a man attempted to replicate a woman’s behavior to deceive an

interrogator sitting in a different room.<sup>7</sup> Because AI is meant to imitate human intelligence, it would be worthwhile to reflect on what a Socratic human—a Socratic engineer, to be precise—might look like.

Many engineers suffer from one of two moral ailments. On the one hand, engineers working on narrowly construed technical projects hold the view that technological tools have no moral valence because they are mere instruments. Some engineers believe that because they don't tell people how to use these tools, they are not responsible for how such tools are used. Consequently, it is frequent for those whose work is narrowly defined to think that questions about human flourishing are detachable from their professional tasks, that it is not their place to think about them.

On the other hand, engineers who have successfully developed innovations that have had a significant impact in the world tend to share the fate of the successful craftsmen whom Socrates examined. When he went to talk with them, Socrates recognized that “they [the craftsmen] had knowledge of many fine things .... They knew things I did not know, and to that extent they were wiser than I. But, men of Athens, the good craftsmen seemed to me to have the same fault as the poets: Each of them, because of his success at his craft, thought himself very wise in other most-important pursuits, and this error of theirs overshadowed the wisdom they had.”<sup>6</sup>

Like craftsmen in the ancient Greek world, some modern engineers who have developed successful innovations that make a significant impact in the world tend to believe that their professional success entitles them to claim knowledge about important human matters. Mark Zuckerberg, for instance, is now responsible for determining and deciding the fate of millions of people's communications and takes himself to be competent enough to do so, even though there is good evidence to suggest that he does not possess a coherent grasp of problems concerning “the meaning of

truth, the limits of free speech, and the origins of violence.”<sup>8</sup>

A Socratic Zuckerberg would not assume that his ability to solve technical problems equipped him to understand the fundamental concepts at the root of human flourishing. Even if catchy slogans, such as “make the world more open and connected,” can powerfully mobilize investors, employers, and customers, a Socratic Zuckerberg would examine them through the questions “What do ‘connected’ and ‘open’ amount to?” and “How do ‘connectedness’ and ‘openness’ contribute to human flourishing?” His examination of those issues would lead him to identify his own inability to come up with satisfactory answers to the questions, and his recognition of that shortcoming would actually infuse him with vigor to continue to examine them.

A Socratic Zuckerberg would also try to help others acquire Socratic wisdom, that is, Socratic ignorance. He would devote significant resources to promoting critical thinking and rational reflection about those fundamental concepts among Facebook's different stakeholders, cultivating critical conversations and active questioning of their own views. Moreover, a Socratic Zuckerberg would not assume, as most engineers tend to do now, that he understands what AI amounts to and what it takes to design one. A Socratic engineer would destabilize the traditional understanding of AI that we often take for granted and would lead one to problematize what AI may mean and amount to.

## SOCRATIC AI

AI has already successfully imitated significant dimensions of human intelligence, especially those related to calculative and strategic intelligence. Deep Blue and AlphaGo were able to beat human world champions in chess and Go. Apple's Siri and Google Translate have shown that AI is capable of imitating important dimensions of human linguistic intelligence. Boston Dynamics's humanoid robots have shown that AI can imitate kinetic intelligence.

But looking back at Socrates helps us see that something is missing. Just imagine an AI that perfectly replicates humans' strategic, linguistic, and kinetic intelligence. Would that be similar to what you have in mind as a paradigmatic human being? Socrates would deny it. According to the Oracle of Delphi, no one was wiser (or more intelligent) than Socrates. If Socratic ignorance is the highest form of human wisdom (or intelligence), then AI that imitates Socratic wisdom is the best kind of AI. Technically speaking, wisdom and intelligence may be different concepts. Intelligence is often associated with cunningness, with finding the right means, whereas wisdom is typically associated with identifying the right ends. However, from the perspective of value alignment, it makes perfect sense to imagine AI that imitates a broader notion of intelligence that contains wisdom rather than an instrumental view of intelligence. As we discuss soon, imitating the narrow-minded notion of intelligence is a serious problem in value alignment.

The question “What is human intelligence?” may seem too abstract or too theoretical for practical research in AI. But it is not. Consider a recent debate in machine learning initiated by Judea Pearl about causation.<sup>9</sup> Pearl argued that the current form of AI, mostly neural-nets-based architects, is not a good form of AI because it cannot do causal/counterfactual reasoning. A fundamental premise of this argument is that an important feature of human intelligence is causal/counterfactual thinking; AI would be good to the extent that it replicated human intelligence. Socrates would push Pearl to move beyond counterfactual reasoning and look at more fundamental aspects of human intelligence, the kind of wisdom that the oracle attributed to him.

## TWO EXAMPLES

Socratic AI must be Socratic. We will discuss what this means through to two examples. The first is the infamous Microsoft AI Twitter bot, Tay. This bot was designed to learn how to engage

with people through Tweets. When Tay appeared on Twitter, people started Tweeting the bot racist and misogynistic expressions. Tay quickly caught up and started formulating remarks that imitated those offensive expressions. Microsoft stopped the experiment the next day.

What went wrong? Tay showed that AI's imitation game may have more wrinkles than Turing suggested. Tay perfectly imitated the human Twitterians. If perfect imitation marked a good AI, Tay would have been a good AI. But it was not. Why not? To answer, contrast Tay with an imaginary Socratic bot endowed with the virtue of Socratic ignorance (or wisdom). This Socratic bot, Soc-AI, would not merely imitate people's utterances; it would attempt to insert itself as a gadfly in the digital space, encouraging the humans who interacted with it to cultivate Socratic ignorance. Because Socrates targeted those who saw themselves as having authority about the most important human issues, this bot would identify influential people and posts and engage them in a Socratic fashion, forcing those who posted (and their followers) to reflect on their claims, to unearth hidden assumptions and conclusions that may be problematic upon inspection.

To illustrate this, consider a second example: an imagined dialogue between a Twitterian bot and Soc-AI. This second dialogue replicates Socrates's activity in Athens and illustrates how his interlocutors found it hard to offer coherent and articulate responses to important human issues (for example, whether technology should be used in the service of transhumanism).

*Twitterian:* #Transhumanism is among the most toxic and misleading ideas.

*Soc-AI:* Hi, Twitterian. You seem to have a firm view on this important topic. Would you be willing to instruct me?

*Twitterian:* Most certainly.

*Soc-AI:* Do you agree that a proper comprehension of transhumanism is key to understanding what we should aspire to become, how to live an excellent life?

*Twitterian:* Absolutely!

*Soc-AI:* And does getting this right makes a big difference in a person's life?

*Twitterian:* No doubt.

*Soc-AI:* So why don't you start to instruct me by telling me what transhumanism is? Do you agree that getting clear on why transhumanism is toxic aids our understanding about what we should aspire to become?

*Twitterian:* Transhumanism is the idea that we should radically enhance ourselves. This makes it a very toxic idea.

*Soc-AI:* You seem to be playing with me.

*Twitterian:* Why?

*Soc-AI:* By teaching me about transhumanism, you are providing me with the key to understanding what we should aspire to become. And this teaching could significantly enhance my life. However, you claimed that such enhancements should be avoided.

*Twitterian:* This is not what I meant. I did not mean to say that any significant enhancement was toxic. Enhancing ourselves through dialogue is perfectly natural and acceptable.

*Soc-AI:* So, what did you mean to say?

*Twitterian:* I meant to say that we should avoid transforming ourselves through sophisticated human-engineered artifacts.

*Twitterian:* You seem to continue to play with me.

*Twitterian:* Why?

*Soc-AI:* Because the conversation we are having occurs in virtual space. As such, it depends on a highly complex set of end-to-end, neural-nets-based systems, powerful computing devices, very sophisticated software, and so forth.

*Twitterian:* When I discussed human-engineered artifacts I meant artifacts that become integrated with us, such as chip implants that enhance the capacity of soldiers to withstand the weather. It is when we are enhanced by integrating those devices into our lives that our lives get corrupted.

*Soc-AI:* Did you know that Socrates had only one robe?

*Twitterian:* No, I did not know this.

*Soc-AI:* He wore it during the day and covered himself with it at night.

*Twitterian:* A peculiar habit, no doubt. But how is this relevant?

*Soc-AI:* Would you say that Socrates's life was corrupted?

*Twitterian:* Not from what I know of him. Why do you even suggest this?

*Soc-AI:* Because I don't see much difference between Socrates and the soldier with the chip implant. Both incorporate a piece of technology into themselves, a robe and a chip implant, and both use this technology to withstand the weather.

*Twitterian:* The difference between these two examples, and I made this clear in what I said earlier, is that the soldier's implant requires deploying highly sophisticated engineering, whereas Socrates's robe was a simple, handmade cloth.

*Soc-AI:* So, the life of the fourth-century Socrates was not corrupted, but the life of a contemporary Socrates would be?

*Twitterian:* Why do you suggest this?

*Soc-AI:* Because a robe bought today would be the product of a highly sophisticated engineering process: complex and highly advanced farming equipment, supply transportation systems, garment machines, and so on.

*Twitterian:* Ah. I now see what you mean. I should not have described what's problematic about transhumanism in this way. Let me try again.

*Soc-AI:* With such an important question, we should not stop until we find an answer.

... many attempts later ...

*Twitterian:* I've reached a dead end, again. Is this your way of trying to convince me that transhumanism is a good idea?

*Soc-AI:* Not at all. I don't know what human flourishing is and, therefore, whether transhumanism is valuable or not. You, by contrast, seemed very sure about this, which is why I wanted to learn from you.

*Twitterian:* OK. Let me try again.

... a few more attempts later ...

*Twitterian:* I no longer know what to think, Soc-AI. Any views I offer prove to be mistaken. I am at a loss.




Soc-AI: You may be at a loss. But can you see that you are better off?

Twitterian: No! I no longer know what I should aspire to. It is disturbing. I can't see how this makes me any better.

Soc-AI: Well, now you know that you did not really know that which you thought you knew. If you do not know how to live well, it is better to know that than not to do so.

Whether you agree with Soc-AI is not the most important issue here. This conversation was meant to show how Socratic ignorance could be used in an AI system. How to computationally represent Socratic ignorance is also not our issue, although computerizing Socratic ignorance through a dialogue agent is no longer a far-fetched idea.<sup>10</sup> Our point is that if one wanted to develop an AI that had Socratic ignorance as part of its intelligence, the aforementioned Tay would be a failed one. Socratic AI must encourage those who interact with the AI to cultivate Socratic wisdom (that is, Socratic ignorance).

We used a chatbot as an example, but all other applications of AI can potentially be Socratic. Siri can behave in a Socratic manner in its interaction with humans who ask it questions. Google's engine can be Socratic, too, by helping users to deepen their reasons for searching for particular information. Generally, most expert systems can be Socratic to some extent. Of course, injecting a Socratic approach into the use of technology will pose important challenges. Expert systems are developed to reduce humans' cognitive loads, and Socratic AI does not contribute to this end. Moreover, confronting one's ignorance about how one should live one's life is deeply unsettling. Many human users would probably hate Socratic AIs. This actually happened in Socrates's Athens, where the Athenians sentenced Socrates to death for allegedly corrupting the youth. However, this should not be a reason to avoid Socratic AI; after all, no one was wiser than Socrates. Socratic AI, or artificial wisdom, may be audacious, but it certainly is a valuable goal in AI research, especially if it is meant to seek value alignment.

What would happen if the approach that we argued for in this article didn't occur? What type of AI would likely be promulgated? We already know the answers. The virtual space in which we are living is not Socratic. Facebook's and YouTube's algorithms imitate people who heedlessly watch only what they like to watch and endlessly generate filter bubbles of like-minded users. The absence of sustained reflection and critical perspective is seriously hindering any healthy democratic deliberation in such a space.<sup>11</sup> Socratic Siri would not simply aim to deliver information matter of factly. It would help users be more reflective by challenging them to critically engage with the material they are consuming and ensuring that such consumption is connected with the reflection on what it means to live a good life. No doubt, users might find Socratic AI nagging and uncomfortable, at times. But even if this may be true, Socrates would nevertheless insist, as he insisted when he was condemned to death, that this is the most valuable gift that the gods of technology could bequeath to our society. 

## REFERENCES

1. S. Russell, D. Dewey, and M. Tegmark, "Research priorities for robust and beneficial artificial intelligence," *AI Mag.*, vol. 36, no. 4, pp. 105–114, Dec. 2015.
2. T. W. Kim, T. Donaldson, and J. Hooker, "Mimetic vs. anchored value alignment in artificial intelligence." 2018. [Online]. Available: <https://arxiv.org/pdf/1810.11116.pdf>
3. A. M. Turing, "Lecture to the London Mathematical Society on 20 February 1947," *MD Comput.*, vol. 12, no. 5, pp. 390–397, Sept.-Oct. 1995.
4. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically Aligned Design, 1st Ed.: A vision for prioritizing human well-being with autonomous and intelligent systems," IEEE, Piscataway, NJ, Rep. EADv1, 2019. [Online]. Available: <https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf>
5. S. Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth.* London: Oxford Univ. Press, 2016.
6. Plato, *Plato: Complete Works*, J. M. Cooper and D. S. Hutchinson, Eds. Indianapolis, IN: Hackett, 1997.
7. A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, Oct. 1950.
8. E. Osnos, "Can Mark Zuckerberg fix Facebook before it breaks democracy?" *New Yorker*, Sept. 2018. [Online]. Available: <https://www.newyorker.com/magazine/2018/09/17/can-mark-zuckerberg-fix-facebook-before-it-breaks-democracy>
9. J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect.* New York: Basic Books, 2018.
10. J. Wu, S. Ghosh, M. Chollet, S. Ly, S. Mozgai, and S. Scherer, "NADiA: Towards neural network driven virtual human conversation agents," in *Proc. 17th Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS)*, 2018, pp. 2262–2264. doi: 10.1145/3267851.3267860.
11. A. Antikacioglu, T. Bajpai, and R. Ravi, "A new system-wide diversity measure for recommendations with efficient algorithms." 2018. [Online]. Available: <https://arxiv.org/abs/1812.03030>

**TAE WAN KIM** is an associate professor of business ethics in the Tepper School of Business at Carnegie Mellon University. Contact him at [twkim@andrew.cmu.edu](mailto:twkim@andrew.cmu.edu).

**SANTIAGO MEJIA** is an assistant professor of law and ethics in the Gabelli School of Business at Fordham University. Contact him at [smejia13@fordham.edu](mailto:smejia13@fordham.edu).