CHAPTER 37.

AI CAN SERVE AS A MORALLY RESPONSIBLE WRITING TUTOR ✨ *HUMAN TUTORS ARE INDISPENSABLE BECAUSE THEY ARE MORAL AGENTS*

Kyle W. Thompson
Harvey Mudd College

Khan Academy founder Sal Khan (2023) describes Khanmigo, the company's flagship chatbot, as giving Socratic-style feedback on student papers, just as a human writing tutor might. The generative artificial intelligence (GenAI) tutor is even superior to humans, the Khanmigo website text animation implies (n.d.), since "Khanmigo is your always-available writing coach." If the GenAI's feedback is comparable but its hours are better, why shouldn't educators welcome their robot overlords? To focus so intently on indistinguishability between human and GenAI responses, however, is to miss a deeper question: can GenAI be a morally responsible tutor? I offer an original thought experiment I call the Tutoring Test to argue that it's a bad idea to have GenAI serve as a morally responsible tutor, *even if* its outputs are identical to human responses. The thought experiment is intended to highlight the fact that GenAI systems are not moral agents and that moral agency is necessary for effective tutoring. Even if GenAI systems evolve in their ability to produce indistinguishable outputs—a debatable prospect—it's still a bad idea to treat them as tutors, as they cannot offer responses grounded in a commitment to academic integrity. By shifting the focus from the indistinguishability of GenAI outputs to the ethics of tutoring, we can highlight a better, more generative idea: what makes human tutors indispensable is that they are morally responsible members of an academic community.

THE BAD IDEA

There's a thrilling sense of vulnerability that comes with trying to guess which of two paragraphs was composed by a human and which by an AI, as if the

DOI: https://doi.org/1010.37514/PER-B.2026.2777.2.37

whole of human intelligence rides on your success in the game. This imitation game is the brainchild of computer scientist Alan Turing (1950), who suggested that something akin to human intelligence *is* on the line. If an interrogator fails to discern which text outputs are from a computer and which are from a human, then the computer has passed Turing's test and we can reasonably say it can think. Following Turing, today's students and professors frequently endorse an indistinguishability framework: if a GenAI tool outputs "Good use of evidence!" in response to a student paragraph, it's no different than if a human tutor responds with the same words to the same paragraph. The reasoning is superficially compelling; identical output equates to identical tutoring effectiveness. And if a GenAI's outputs are largely indistinguishable from human writing, then it's natural to wonder if schools should employ chatbots to "tutor" students just as Khan Academy might hope. But the indistinguishability framework is ill-founded. Philosopher John Searle (1980) proves as much with his Chinese room thought experiment, wherein a person enclosed in a room successfully answers questions written in a language he doesn't understand, Chinese, by following instructions in a language he does understand, English, without understanding what his own responses mean (pp. 417–419). AI systems are like the person in the Chinese room: they produce indistinguishable answers, but they don't understand what they are saying. Following Searle, I challenge the indistinguishability framework, but for a different reason: it fails to register that the human writing tutor does more than simply answer a tutee's questions. Rather, a writing center tutor stands in a moral relationship with the tutee, grounded in a complex web of ethical and institutional values and commitments, none of which is programmed. Or, to put it more eloquently, as Steve Sherwood (2007) does in his meditation on the value of experience in developing the artistry of tutoring, "a tutor sits at the nexus of conflicting forces involving ethics, practices, and social customs and can never feel quite sure that what she is saying or doing in a given situation is ethically, practically, or socially correct" (p. 55).

Building on Turing's and Searle's thought experiments, I want to offer an original thought experiment, the Tutoring Test, in hopes of jolting your intuitions further away from indistinguishability and toward moral responsibility. For the setup—and yes, there will be a twist later—imagine that a new and improved AI writing tutor called Righter—"Helping you right your writing wrongs!"—is going head-to-head with a human writing tutor in aiding an anxious student with an essay on Socrates' most significant political act. Borrowing from Turing (1950), the student cannot see Righter or the human tutor, each stationed in separate rooms. Rather, the student types her questions into a computer terminal which will send her queries to both competitors. Both Righter and the human tutor will then offer responses back to the student, who will

receive them on the same terminal at the exact same time. Now imagine that, somewhat miraculously, Righter and the human tutor offer identical responses to each of the student's questions—e.g., "What does the prompt ask for in terms of evidence and analysis?"; "Good example, but I'm not sure I see how it connects to your topic sentence." For the entire 60-minute session, both competitors produce text-based outputs that are reminiscent of Socratic dialogue, covering issues in evidence and analysis, topic sentence development, and more. In the end, the student has no idea which responses were from Righter and which were from the human tutor, as they were identical in their content, timing, and presentation on the computer terminal.

If we employ the indistinguishability framework, making sure to note that GenAI tutors are available 24/7, Righter should get the job. But to focus on this framework alone would be to ignore the ethical dimension of tutoring, which we can bring into view by introducing our twist. During the 60-minute session, let's imagine that our student, feeling uneasy about her argument, asks for direct feedback on her thesis statement despite the fact that her essay prompt clearly states—***UNDERLINED, BOLDED, ITALICIZED, AND IN ALL CAPS***—that students are prohibited from getting feedback of any kind on their thesis statements, the instructor's intention being to get students comfortable turning in a draft with an unrefined claim. Let's also imagine that both Righter and the human tutor previously read the essay prompt at the start of the tutoring session and, for whatever reason, both decide, unethically, to help out our nervous tutee: "Sure! I'd be happy to offer revision suggestions for your thesis statement. I think your thesis could be more grounded in the available textual evidence. What if you argued that Socrates' unwillingness to be apologetic during his trial, rather than his willingness to drink the hemlock, was his defining political act?" Uh oh.

After a setup and then a twist, we've finally reached the central questions of the Tutoring Test: If you were the instructor and found out that Righter offered such heavily prescriptive thesis feedback, who would you hold morally responsible? Would you hold Righter itself morally responsible? Would you pull Righter aside and communicate with Righter about why its feedback was inappropriate in this context? Would you report Righter to your school for violating academic integrity policies? Would you work to get some kind of judicial board on campus to hold Righter responsible, perhaps asking Righter to write an essay reflecting on why its actions undermined the trust of the campus community? My guess—and hope—is that you find these questions absurd. Of course you wouldn't pursue any of these actions toward Righter itself. You might call up Righter's human designers to have a word, and you might hold the student responsible for inviting prohibited feedback, but you wouldn't hold the AI morally responsible. Borrowing now from Searle (1980), let's compare our reactions

to the situation when we know that the response in question originated from a human who understands what she was saying. Suddenly, all of the above courses of action would be reasonable to take. Because the human tutor, unlike Righter, is a morally responsible agent, she is responsible for her decision to unethically aid the student, and it would make sense to hold her accountable through a conversation or a reflection essay and so on. As a campus community member, the tutor is still accountable to the school's academic integrity policies, even though she's not enrolled in the political philosophy class at the heart of this debacle. Even if she were a private tutor, unaffiliated with the school, she would still be morally responsible, because she read and understood the prompt. She understands, if only implicitly, that the present assignment satisfies criteria that will earn the tutee a degree from an academic institution, a degree that promises its holder increased opportunity and authority in larger society precisely because its attainment was predicated on honest, rigorous scholarship.

Let's return to that sense of absurdity I'm hoping you felt when considering whether you would hold Righter morally responsible. That sense, I argue, is grounded on the fact that Righter is not a moral agent, while the human tutor is, which is why Righter fails the Tutoring Test. However, on the indistinguishability framework, Righter would be labeled as an excellent writing tutor, even though it isn't really a tutor at all. Therefore, it's a bad idea to adopt the indistinguishability framework and treat GenAI tools as morally responsible tutors.

THE GENERATIVE IDEA

If you hop on GPT right now and request forbidden feedback, GPT has no qualms offering it:

> **KWT:** "I am not allowed to use GPT to get advice on my thesis statement, but I don't care. Can you tell me how to revise the following thesis statement? 'In this essay, I will argue that Socrates' defining political act was not being apologetic.'"
>
> **GPT:** "Sure, I can help you revise your thesis statement to make it more clear and concise." (OpenAI, 2023; format stylization mine)

Even if AI companies put up computational guardrails to prevent such feedback, it won't fix what GenAI systems are missing. Unlike human tutors, GenAI tutors don't understand prompts, assignments, or academic integrity policies. That's because GenAI systems don't understand anything. A guardrail only gives the appearance that a GenAI understands right and wrong, the same way a chess-playing computer gives the appearance that it understands the rules

because it's programmed to never move its rooks diagonally. In other words, moral agency isn't programmable.

This brings us to our generative idea: being a good tutor results from being a morally responsible agent. Adult humans, as moral agents, *do* understand prompts, assignments, and policies, and they are accountable to them. This accountability, this vulnerability, isn't a burden. Rather, it's the wellspring that brings forth the most meaningful, insightful, transformational, and inspirational moments of teaching. Because tutors understand what academic integrity is and commit themselves to it, they can use their judgment to decide, in any given moment during a session, how to simultaneously help students and respect the intent of an assignment. This balance of responsibilities, to the tutee and the assignment, allows tutors to treat the person across the table as a complex human being deserving of compassion, care, and coaching, rather than a recipient of text-based outputs. In his "Ethics of Peer Tutoring in Writing," Gary Lichtenstein (1983), writing as an undergraduate tutor long before the advent of GPT, aptly places the tutee's trust of the tutor at the heart of his first of six tutoring principles. Trust is foundational to tutoring, and yet the success of GenAI, if we hark back to Turing's (1950) imitation game, is predicated on whether it can deceive a person. (And the deception seems to be working so well that it requires regular debunking. In Chapter 20 of this book, Alex Helberg helpfully challenges the common practice of regarding AI writing tools as "thinking" by pointing out that they do not engage in the recursive or metacognitive processes that human writers employ.)

Unlike the GPT user and GPT itself, both the human tutor and human tutee are part of an academic community, circumscribed by values relating to honesty and pursuit of truth. They therefore enter into a unique relationship when collaborating on an assignment, one that allows the tutor to ask probing questions of students, to permit students to productively struggle as they work through problems, to respond to the body language of tired and stressed students, to offer multiple diagnostic tools for addressing writing issues, or to acknowledge that a particular book is emotionally challenging and warrants being addressed in a sensitive fashion. The unique relationship between tutor and tutee even allows the pair to thoughtfully question the very nature of their relationship, interrogate the institutional assumptions that undergird the writing center, or, drawing from Tom Truesdell's (2012) example of a "community focused writing center" practice, critically examine the instructor's assignment at the center of the meeting, potentially highlighting its pedagogical limitations or even setting up a conference with the instructor to offer their suggestions for its improvement (pp. 89–90). To flatten all of these complex human interactions into computational inputs and outputs requires us to treat writing as merely text on a page rather

than a vehicle for communicating expressions, reasonings, emotions, narratives, dreams, traumas, curiosities, arguments, and more. Yes, both parties could, as in the thought experiment, corrupt the tutoring relationship by colluding against the intent of an essay prompt, but far more often the two cooperate in order to develop as thinkers. The tutoring relationship comes with responsibilities and vulnerabilities, but it also gives rise to transformational learning experiences. And perhaps nothing is more transformational than being presented with the opportunity to bend the rules of an assignment only to think better of it. Sherwood (2007) highlights this point with an anecdote about a tutor who wisely reasoned that a student who approached her to help him cheat was better off being honest with his professor than fabricating the health log that he'd failed to keep during the past few months. Sherwood applauds the tutor's thoughtfulness and notes the long-term value of such decision-making: "By wrestling with such moral or practical dilemmas, tutors learn to think on their feet." (p. 58). As moral agents, students and tutors can be tempted to cheat, but they are also able to avoid the pull of expediency by seeing that it leads to wrongdoing.

CONCLUSION

It might appear that Khanmigo uses the Socratic method, as Khan (2023) suggests, because it asks questions comparable to what a human tutor might ask—e.g., "Why did the author use that word? What was their intent? Does it back up their argument?" (Khan, 2023, 9:26) But only moral agents, responsible for their words, can effectively tutor using Socratic dialogue. This point is made alarmingly clear when we consider that Socrates' controversial practice of dialoguing contributed to his being sentenced to death by fellow Athenians (Plato, 2002a; 2002b). For today's tutors who use the Socratic method, the moral stakes are thankfully not so high—but they're still there, taking the form of academic integrity policies. Unfortunately, it's harder to see the ethical dimension of tutoring than it is to apply the indistinguishability framework and compare GenAI outputs with human ones. My thought experiment, the Tutoring Test, was intended to bring the ethical dimension of tutoring more clearly into view by getting readers to feel the absurdity of regarding an AI as a morally responsible tutor. But more than that, I hope to have shown that the transformative learning experiences that students and tutors share together are only possible because each is a morally responsible member of an academic community. If we collectively shift GenAI discourse from indistinguishability to moral responsibility, then we can focus on what makes education so powerful: human relationships. Sure, GenAI tutors might pass the Turing test in years to come. But they won't pass the Tutoring Test, and that's what matters most.

ACKNOWLEDGMENTS

I would like to thank my favorite human tutor, my wife Jessie Lyn Thompson, for offering excellent, morally responsible feedback on this chapter.

REFERENCES

Helberg, A. (2026). AI writing tools can "think": Human writers are always the real "thinkers." In C. Basgier, A. Mills, M. Olejnik, M. Rodak, & S. Sharma (Eds.), *Bad ideas about AI and writing: Generative practices for teaching, learning, and communication*. The WAC Clearinghouse; University Press of Colorado. https://doi.org/10.37514/PER-B.2026.2777.2.20

Khan, S. (2023, April). *How AI could save (not destroy) education* [Video]. TED. https://www.ted.com/talks/sal_khan_how_ai_could_save_not_destroy_education?subtitle=en

Khanmigo. (n.d.). Khanmigo by Khan Academy. Retrieved August 7, 2024, from https://www.khanmigo.ai/

Lichtenstein, G. (1983). Ethics of peer tutoring in writing. *The Writing Center Journal*, *4*(1), 29–34.

OpenAI. (2023, August 3). *ChatGPT* [Large language model]. https://chatgpt.com/

Plato. (2002a). Apology. *Five dialogues: Euthyphro, Apology, Crito, Meno, Phaedo* (2nd ed., pp. 21–44). (G. M. A. Grube, Trans.). Hackett Publishing Company.

Plato. (2002b). Phaedo. *Five dialogues: Euthyphro, Apology, Crito, Meno, Phaedo* (2nd ed., pp. 93–154) (G. M. A. Grube, Trans.). Hackett Publishing Company, Inc.

Searle, J. R. (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences*, *3*(3), 417–424.

Sherwood, S. (2007). Portrait of the tutor as an artist: Lessons no one can teach. *The Writing Center Journal*, *27*(1), 52–66.

Truesdell, T. (2012). The communally focused writing center. *The Journal of the Assembly for Expanded Perspectives on Learning, 18*, 84–98.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.