Where does human thinking end and AI begin? An AI authorship protocol aims to show the difference

Eli Alshanetsky, Assistant Professor of Philosophy, Temple University

Published: October 30, 2025 8:26am EDT

If students can't demonstrate their thinking, how can professors know whether they are learning?
SDI Productions via Getty Images

The latest generation of artificial intelligence models is sharper and smoother, producing polished text with fewer errors and hallucinations. As a philosophy professor, I have a growing fear: When a polished essay no longer shows that a student did the thinking, the grade above it becomes hollow – and so does the diploma.

The problem doesn't stop in the classroom. In fields such as law, medicine and journalism, trust depends on knowing that human judgment guided the work. A patient, for instance, expects a doctor's prescription to reflect an expert's thought and training.

AI products can now be used to support people's decisions. But even when AI's role in doing that type of work is small, you can't be sure whether the professional drove the process or merely wrote a few prompts to do the job. What dissolves in this situation is accountability – the sense that institutions and individuals can answer for what they certify. And this comes at a time when public trust in civic institutions is already fraying.

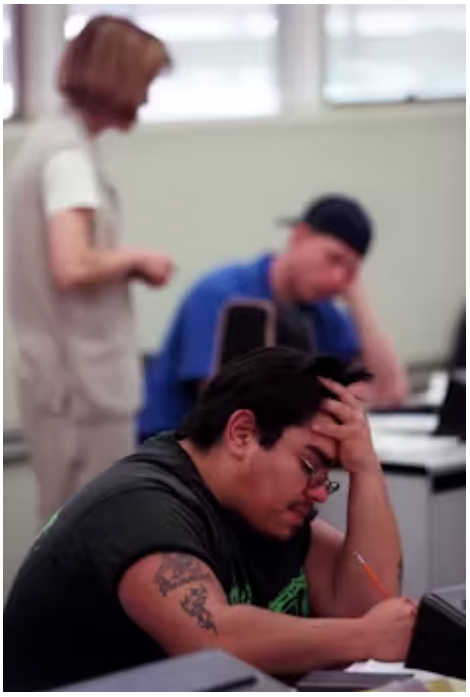
I see education as the proving ground for a new challenge: learning to work with AI while preserving the integrity and visibility of human thinking. Crack the problem here, and a blueprint could emerge for other fields where trust depends on knowing that decisions still come from people. In my own classes, we're testing an authorship protocol to ensure student writing stays connected to their thinking, even with AI in the loop.

When learning breaks down

The core exchange between teacher and student is under strain. A recent MIT study found that students using large language models to help with essays felt less ownership of their work and did worse on key writing-related measures.

Students still want to learn, but many feel defeated. They may ask: "Why think through it myself when AI can just tell me?" Teachers worry their feedback no longer lands. As one Columbia University sophomore told The New Yorker after turning in her AI-assisted essay: "If they don't like it, it wasn't me who wrote it, you know?"

Universities are scrambling. Some instructors are trying to make assignments "AI-proof," switching to personal reflections or requiring students to include their prompts and process. Over the past two years, I've tried versions of these in my own classes, even asking students to invent new formats. But AI can mimic almost any task or style.



In-class assignments on paper can get around student dependence on AI chatbots. But 'blue book' exams emphasize performance under pressure and may not be good for scenarios where students need to develop their own original thinking.

Robert Gauthier/Los Angeles Times via Getty Images

Understandably, others now call for a return to what are being dubbed "medieval standards": in-class test-taking with "blue books" and oral exams. Yet those mostly reward speed under pressure, not reflection. And if students use AI outside class for assignments, teachers will simply lower the bar for quality, much as they did when smartphones and social media began to erode sustained reading and attention.

Many institutions resort to sweeping bans or hand the problem to ed-tech firms, whose detectors log every keystroke and replay drafts like movies. Teachers sift through forensic timelines; students feel surveilled. Too useful to ban, AI slips underground like contraband.

The challenge isn't that AI makes strong arguments available; books and peers do that, too. What's different is that AI seeps into the environment, constantly whispering suggestions into the student's ear. Whether the student merely echoes these or works them into their own reasoning is crucial, but teachers cannot assess that after the fact. A strong paper may hide dependence, while a weak one may reflect real struggle.

Meanwhile, other signatures of a students' reasoning – awkward phrasings that improve over the course of a paper, the quality of citations, general fluency of the writing – are obscured by AI as well.

Restoring the link between process and product

Though many would happily skip the effort of thinking for themselves, it's what makes learning durable and prepares students to become responsible professionals and leaders. Even if handing control to AI were desirable, it can't be held accountable, and its makers don't want that role. The only option as I see it is to protect the link between a student's reasoning and the work that builds it.

Imagine a classroom platform where teachers set the rules for each assignment, choosing how AI can be used. A philosophy essay might run in AI-free mode – students write in a window that disables copy-paste and external AI calls but still lets them save drafts. A coding project might allow AI assistance but pause before submission to ask the student brief questions about how their code works. When the work is sent to the teacher, the system issues a secure receipt – a digital tag, like a sealed exam envelope – confirming that it was produced under those specified conditions.

This isn't detection: no algorithm scanning for AI markers. And it isn't surveillance: no keystroke logging or draft spying. The assignment's AI terms are built into the submission process. Work that doesn't meet those conditions simply won't go through, like when a platform rejects an unsupported file type.

In my lab at Temple University, we're piloting this approach by using the authorship protocol I've developed. In the main authorship check mode, an AI assistant poses brief, conversational questions that draw students back into their thinking: "Could you restate your main point more clearly?" or "Is there a better example that shows the same idea?" Their short, in-the-moment responses and edits allow the system to measure how well their reasoning and final draft align.

The prompts adapt in real time to each student's writing, with the intent of making the cost of cheating higher than the effort of thinking. The goal isn't to grade or replace teachers but to reconnect the work students turn in with the reasoning that produced it. For teachers, this restores confidence that their feedback lands on a student's actual reasoning. For students, it builds metacognitive awareness, helping them see when they're genuinely thinking and when they're merely offloading.

I believe teachers and researchers should be able to design their own authorship checks, each issuing a secure tag that certifies the work passed through their chosen process, one that institutions can then decide to trust and adopt.

How humans and intelligent machines interact

There are related efforts underway outside education. In publishing, certification efforts already experiment with "human-written" stamps. Yet without reliable verification, such labels collapse into marketing claims. What needs to be verified isn't keystrokes but how people engage with their work.

That shifts the question to cognitive authorship: not whether or how much AI was used, but how its integration affects ownership and reflection. As one doctor recently observed, learning how to deploy AI in the medical field will require a science of its own. The same holds for any field that depends on human judgment.

I see this protocol acting as an interaction layer with verification tags that travel with the work wherever it goes, like email moving between providers. It would complement technical standards for verifying digital identity and content provenance that already exist. The key difference is existing protocols certify the artifact, not the human judgment behind it.

Without giving professions control over how AI is used and ensuring the place of human judgment in AI-assisted work, AI technology risks dissolving the trust on which professions and civic institutions depend. AI is not just a tool; it is a cognitive environment reshaping how we think. To inhabit this environment on our own terms, we must build open systems that keep human judgment at the center.

Eli Alshanetsky does not work for, consult, own shares in or receive funding from any company or organization that would benefit from this article, and has disclosed no relevant affiliations beyond their academic appointment.