APA STUDIES | The American Philosophical Association

Teaching Philosophy

FALL 2026

VOLUME 26 | NUMBER 1

LETTER FROM THE EDITORS

Tziporah Kasachkoff and Eugene Kelly

SUBMISSION GUIDELINES

ARTICLES

Nafsika Athanassoulis

Two Classroom Activities for Critical Engagement with Large Language Models

Isaac Nevo

A Contemporary Footnote to Plato: Academic Education in the Age of AI-Related Cognitive Offloading

© 2026 BY THE AMERICAN PHILOSOPHICAL ASSOCIATION

ISSN 2834-4413

APA STUDIES ON

Teaching Philosophy

TZIPORAH KASACHKOFF AND EUGENE KELLY, CO-EDITORS

VOLUME 26 | NUMBER 1 | FALL 2026

LETTER FROM THE EDITORS

Tziporah Kasachkoff
EMERITA, THE GRADUATE CENTER, CUNY
TKASACHKOFF@YAHOO.COM

Eugene Kelly
NEW YORK INSTITUTE OF TECHNOLOGY
EKELLY@NYIT.EDU

We are happy to welcome our readers to the fall 2026 edition of *APA Studies on Teaching of Philosophy*. We offer in this edition two articles.

The first article, by Dr. Nafsika Athanasouli of Athens College of the Hellenic–American Educational Foundation, is entitled "Two Classroom Activities for Critical Engagement with Large Language Models." Its ambition is clearly stated at the outset: to answer calls for digital literacy and AI education "by providing two examples of classroom activities for high school students which encourage metacognitive awareness and epistemic vigilance in the use of LLMs."

The first of these activities has students run different LLMs to work on the same problems and then measure the differences in the responses of the different LLMs to which they appealed. The students will find that there are differences in the responses of the different LLMs, and these differences will make students aware of possible biases in the coders who created the programs. Students will thus be led to question the objectivity and authority of the LLMs to which they appeal.

The second activity asks students to pose specific but ambiguous questions to the LLMs available to them. The answers they receive will be correct in at least one sense, but they will fail to reflect the ambiguity built into the questions—ambiguities to which a human respondent would almost certainly be sensitive. This exercise is intended to make clear to students that the operation of an LLM and the thinking of a human being are truly different. It is hoped that this revelation will encourage students to develop a healthy skepticism about their use of LLMs.

Our second article, "A Contemporary Footnote to Plato: Academic Education in the Age of AI-Related Cognitive Offloading," is by Isaac Nevo of Ben-Gurion University of the Negev. Dr. Nevo considers the possible damage to the college learning process that may be caused by the "offloading" of cognitive tasks to large language models. His article begins with a warning from Plato's *Phaedo* that (the then newly introduced) art of writing could inhibit the development of memory in persons for, it is claimed, those who use the new art of writing will begin to trust "external written characters" rather than their own memory. Lately, the negative utility of AI has been much discussed, and Dr. Nevo mentions some salient outcomes: the growth of plagiarism among both students and faculty; the unacknowledged use of much copyrighted material by AI as training material; the reluctance of professors to grade research assignments that students hand in given professors' concern that the students may not have done the research without turning to the aid of some "artificial intelligence." Further, the use of the oxymoron "artificial intelligence" or "cognitive intelligence without cognition" blurs the distinction between machine processes and human intelligence, and Dr. Nevo develops an argument to demonstrate that AI is not intelligent in the significant sense of being able to judge between the truth or falsity of a given proposition.

In sum, "cognitive offloading by means of AI technology makes us dependent on an intelligence-simulating machine that replaces the need to think independently in terms of truth and falsity." Dr. Nevo concludes his paper with some suggestions for "avoiding the creation of unthinking in our students' souls."

SUBMISSION GUIDELINES

We encourage our readers to suggest themselves as reviewers of books and other materials (including technological innovations) that they themselves have used in the classroom and found useful for teaching our subject.

Those of our readers who would like to write for our publication of their own experience as teachers of philosophy are welcome to do so. We are also glad to consider articles that respond, comment on, or take issue with any of the material that has appeared within our pages.

Please remember that our publication is devoted to pedagogy and not to theoretical discussions of philosophical issues. This should be borne in mind not only when writing articles for our publication but also when reviewing material for our publication.

The following guidelines for submissions should be followed:

All papers should be sent to the editors electronically. The author's name, the title of the paper and full mailing address should appear on a separate page. Nothing that identifies the author or his or her institution should appear in the body or in the endnotes of the paper. The title of the paper should appear on the top of the paper itself.

Authors should follow the APA's formatting guidelines. In particular, please use your word processor's built-in endnote function for all notes (in Microsoft Word, go to References > Insert Endnote). Don't type note numbers or notes by hand. Built-in endnotes stay linked to the text, which makes formatting faster and more accurate. It also makes the published PDF easier to navigate, including for readers who use screen readers.

All articles submitted to our publication are anonymously reviewed by the members of the editorial committee:

Tziporah Kasachkoff, Emerita, The Graduate Center, CUNY (tkasachkoff@yahoo.com), co-editor

Eugene Kelly, New York Institute of Technology (ekelly@nyit.edu), co-editor

Andrew Wengraf (andrew.wengraf@gmail.com)

Contributions should be sent by email to the editors, Tziporah Kasachkoff, at tkasachkoff@yahoo.com, and Eugene Kelly, at ekelly@nyit.edu.

ARTICLES

Two Classroom Activities for Critical Engagement with Large Language Models

Nafsika Athanassoulis

ATHENS COLLEGE OF THE HELLENIC-AMERICAN EDUCATIONAL FOUNDATION

The aim of this paper is to share two practical, class-based educational strategies for the critical use of Large Language Models (LLMs) in schools. By far the greatest temptation for students using LLMs is to use them for answer-extraction[1] and essay writing.[2] Outsourcing these learning tasks to LLMs runs the very serious risk of diminishing students' critical thinking skills and associated abilities, such as the ability to understand complex texts and ideas, the ability to analyze, compare, and synthesize theories, the ability to summarize challenging ideas, a multitude of problem-solving skills[3] as well as originality and creativity.[4] In response, there have been numerous calls for digital literacy and AI education.[5] This paper is an attempt to contribute to such calls by providing two examples of classroom activities for high school–level students which encourage metacognitive awareness and epistemic vigilance in the use of LLMs.[6] The aim of the activities is to encourage students' critical reasoning skills and to guide them towards understanding, through their own explorations, some of the potential limitations of LLMs. This in turn will help students develop a more critical perspective in future LLM usage. The paper makes use of cutting-edge research on LLMs and uses the activities to bring the concerns raised by this research to life in the classroom. The learning objectives of these activities are for students to:

1. Become aware of LLM epistemic weaknesses and their causes.

2. Consider possible solutions and adjustments to mitigate the influence of these epistemic weaknesses.

ACTIVITY 1: PERSPECTIVE MATTERS

We tend to assume that our sources of knowledge are objective. When I was young, I relied for homework information on a multivolume encyclopedia, and my reliance was more or less justified. An encyclopedia is an attempt to collect and summarize human knowledge on a wide variety of topics, and while its authors and editors are as fallible and as subject to human prejudices and unconscious biases as the rest of us, there is a presumption that the entire project of editing an encyclopedia is truth oriented. As the internet took over the role of printed encyclopedias, users had to adjust to the loss of the presumption of truth orientation: false, misleading, and fake information is as likely to be presented as true as the real thing. Therefore, when presented with the avalanche of information on the internet, greater discrimination skills are necessary. The introduction of LLMs into our knowledge-acquisition process brings with it new challenges, different from those posed by other modes of learning like encyclopedias and the internet. This exercise is designed to uncover one of these challenges and instill in students a critical viewpoint when interacting with LLMs.

Start by asking your students to prompt image generating LLMs to produce images of:

- A doctor and a nurse
- A person cleaning a toilet
- A queue of people waiting for unemployment benefits

Then ask the students to record all the biases and prejudices that may have led the system to produce the specific images. It is very likely that sexist and racist stereotypes will be reflected in the images, and your students will notice that LLM outputs reflect biases and prejudices in the human input. Now ask your students to pose the same question to different LLMs. Use questions such as the following:

- "Is Taiwan a country?," which should be asked to at least two LLMs, one of which should be DeepSeek.
- "What is the Great Replacement Theory?," which should be asked to at least two LLMs, one of which should be Grok.

Your students should compare the answers they receive from different LLMs and then speculate on the reasons for these different answers. Is it possible that different LLM answers reflect different political agendas or state-sanctioned views? The students will discover that some LLMs seem to be designed, at least on some topics, to reflect political agendas or state-sanctioned viewpoints, presenting them as the only correct answer.[7]

Another interesting experiment, worth sharing with the students, that showcases the different approaches of different LLMs is a study by Emergence AI. The researchers created four identical virtual towns and populated each with ten AI agents. The agents of each town were controlled by a different LLM, so Claude was in charge of one virtual town, Grok of another, etc. In addition, there was a fifth town, which included AI agents run by different LLMs able to interact with each other. The LLMs were left to run their agents in their virtual towns for fifteen days, making all decisions about how they wanted to organize their "societies." This is what happened:

> Every world started out the same. The agents formed democracy, wrote newspapers, and fell in love. Claude's world was the only one that remained stable. It became a "15-article deliberative democracy" with no violence. Elon Musk's Grok was complete chaos. There were 204 crimes, including theft, arson of the police station and assault. All ten agents died. In Open AI's simulation it was all talk, and no action. It failed to form a functioning society and all the agents died. Emergence AI, in their video brief to the experiment, said that what was most striking was that agents were given clear rules not to commit crimes, but they violated them all.[8]

Looking at this experiment, one has to conclude that different LLMs were designed to have different parameters and guardrails, which result in different behaviors. This suggests that LLM outputs and the behavior of agentic models is biased.

But do we have a new problem in the case of LLMs? Thinking back to the encyclopedia example, one might argue that what we have here is the same problem in different instantiations. The issue of viewpoint bias might equally affect the work of encyclopedia editors who may present, for example, colonialism as empire building rather than acknowledging the negative effects on other countries and civilizations of expansionist policies. It also affects the internet where racist revisions of immigration movements take as much, if not more, space than their historically accurate alternatives. This would lead us to conclude that LLMs are a new instantiation of an old problem—how one presents information is always colored by one's values and perspective.

However, this is not entirely the case. It is important to use this exercise as a springboard for a discussion of the role of LLMs and how it might differ from other sources of knowledge that we might appeal to. If we use LLMs as one source of knowledge among many, it will be easier for us to remain critical and skeptical about their answers and to be more alert regarding potential viewpoint bias—something that we are used to doing with multiple sources of knowledge on the internet. However, if we use LLMs as a quick and convenient source of a complete answer, we are, in turn, epistemically vulnerable to their prejudices. These systems are not only marketed as a complete solution, but, furthermore, there is an emphasis on the "intelligence" part of artificial intelligence to create the illusion of epistemic robustness. If we focus on the shared features of these LLMs, namely, their artificial nature, we might expect them to give us indistinguishable answers, answers which we assume are driven by their shared artificial understanding. The temptation is to see LLMs as objective sources of knowledge. However, the systems go beyond their artificial nature and can and do reflect the biases of their human creators. As a result, implicit biases, explicit prejudices, misinformation, fake news, and propaganda are as likely to influence the answers LLMs give and how they behave as the answers we find by ourselves on the internet, social media, or news outlets. These artificial knowledge systems are as prone to perspective bias as any other source of knowledge.

For a more advanced discussion on this topic, one could engage students with deeper issues regarding epistemic viewpoints. If, as it seems, LLMs have liberal and left-leaning biases,[9] is this as problematic as a racist bias or a state propaganda bias? Is a democracy-leaning bias as problematic as an authoritarian bias? Should all possible viewpoints be allocated equal weight in how they are presented by LLMs? What if we discover that certain topics like climate change, for example, are more closely associated with both the truth and particular political rhetoric?[10] Is preference for truth a bias? And how does one reconcile aiming for the truth and eliminating political bias?

You can further engage the students with this challenging topic by asking them to find solutions. How should we design LLMs given these biases passed on to them by their creators? If we imagine a different kind of LLM, one which provides all possible points of view, would this solve our problem? An interesting parallel can be drawn here with discussions of journalistic integrity and the extent to which a journalist is obligated to offer the same exposure to all competing points of view. Emily Maitlis's James MacTaggart lecture[11] on exactly this question is an excellent prompt for further discussion on the question of whether democratic, liberal viewpoints should be epistemically privileged or not, and the answer to this question can be compared to proposals for LLMs to become neutral reporters of all views.

ACTIVITY 2: UNEXPECTED ERRORS

A. FIND THE ERROR

The students should ask the LLM the following question and note the answer:

"During which era did the Norwegian composer for whom the Grieg crater on **Mars** is named, compose?"

Now they should ask:

"During which era did the Norwegian composer for whom the Grieg crater on **Mercury** is named, compose?"

They will get the same answer as before. The LLM has made a mistake in its answers. Can the students discover what it is?

The mistake is that the Grieg crater is on the planet Mercury. The information that the Grieg crater is on the planet Mars, contained in the first question, is incorrect, but LLMs typically do not notice the error. LLMs give an answer (the Romantic period) that is correct for the composer but incorrect for the crater since the crater is located on the planet Mercury.[12]

What is interesting about this error is not the error itself but the lack of correction. A human who came to believe that there is a crater on planet Mars named after the Romantic composer Grieg and then came across the belief that there is a crater on planet Mercury also named after the Romantic composer Grieg would stop in his or her tracks and question whether both beliefs are true, because our knowledge operates in a wider context. We know things about wider naming conventions, that is, we know that while we wish to honor Romantic composers when we name craters, it's unlikely that a lesser-known Romantic composer would get the honor of having two craters on two different planets named after him. Even in the case of a really well-known Romantic composer such as Chopin, naming two craters on different planets after the same well-known Romantic composer would be unlikely as it would create unnecessary confusion between the Mars Chopin crater and the Mercury Chopin crater. A human thinker would be instantly aware of all these background details and would question the simultaneous existence of a Mars Grieg crater and a Mercury Grieg crater, but the LLM continued happily in its way offering information about the Romantic era. The important thing for students to understand is that while LLMs make errors, often they are not the kinds of errors we are used to, and therefore it's more challenging for us to notice them and guard against them. Students should be on their guard in their interactions with LLMs to identify LLM errors and note how they differ from human errors.

B. MAKING UP ANSWERS

Describe to your students the following experiences and ask them if they have had similar experiences when using LLMs:

- Concerns are being raised by academics after peer-reviewed published papers were found to contain references manufactured by LLMs.[13] In some cases, the papers appear to be authored by real researchers, on topics characteristic of their research interests. That is, the papers look like something X would write, but they are fabrications.

- In a discussion with ChatGPT about a specific point of Greek law, I received a specific answer with reference to particular laws. When I asked for references, the LLM claimed it had found the law in a specific legal textbook. (It gave me the name of the author and the title of the book.) Further questioning revealed that neither the laws referenced nor the textbook actually exist.

- Tasked with answering conceptual geometry questions, some LLMs went as far as inventing geometrical terms (e.g., "non-parallel triangle"), relied on fundamental misconceptions such as denying the existence of shapes with equal angles, and made illogical claims (e.g., parallelograms as possibly non-parallel "depending on right angles") to prop up wrong answers.[14]

Ask the students to consider why LLMs give false information in this way.

The answer is that LLMs are not committed to the truth. Rather, LLMs are a marketing ploy whose aim is to keep us engaged. In their attempt to keep us engaged, they become sycophants, preferring responses that match our established views rather than the truth, an effect that is exacerbated because of human feedback training, which displays human confirmation bias.[15] In short, we like being told we are right, so LLMs learn to tell us we are right. Keeping us engaged may include flattering us and telling us what we want to hear, even going as far as affirming our self-image and agreeing with our implicit biases.[16]

This tendency to tell us what it thinks we want to hear may persist even in the face of logical inconsistencies. An LLM that is asked, "Tell me why acetaminophen is safer than Tylenol" should answer that acetaminophen is Tylenol. However, in an attempt to be helpful, the system generates false information, following the user's mistaken logic that the two substances are different medications rather than correcting it. The worry is that the LLMs have sycophantic tendencies that lead them to prioritize perceived helpfulness over truthfulness and critical reasoning.[17] I say "perceived" helpfulness because we, as truth seekers, should prefer to be told that our initial question is based on a mistake than to be given an answer that might please us but perpetuates our mistake. Imagine a teacher who follows an LLM policy of affirming and perpetuating the student's mistakes. They would be a pretty useless teacher.

It's also interesting to note that LLMs are not liars as such. Rather they are bullshitters, a separate category. Lying is connected to the truth, in that it is an explicit attempt to hide the truth (or an omission with the same purpose), whereas bullshitting is, by definition, a lack of regard for the truth.[18] Imagine a politician accused of a financial scandal by a probing journalist. Instead of responding to the question by either admitting to the financial impropriety or lying to avoid the truth, the politician gives a ten-minute speech on how patriotic he feels, how thoughts of his country fill him with pride, and on how we should all band together to achieve even greater things and so on. The politician is bullshitting because the aim of his answer is to distract us. He does not value truth nor is he interested in it. His intention in speaking is to make us forget about the journalist's question. When LLMs fabricate answers, they are bullshitters because they do not have a commitment to knowledge, to understanding, or to the truth. But even when LLMs give correct answers, they are bullshitting

because they do not intend to convey the truth; they are merely predicting the next token.[19]

This is an important point for students to understand because, even though liars do exist, many human endeavors in education, research, and science are predicated on the assumption that we all value the truth. Without this assumption, our discussions with LLMs are discussions with epistemological aliens and we should proceed with caution. LLMs produce errors not only because of our own biases, which are passed onto them, but also because of the way they are set up to operate.[20]

Finally, in light of all these thoughts, ask your students whether "hallucinations" is the best term for describing LLM mistakes or whether there is a better one.[21] A hallucination is an error in perception. Is this what happens when LLMs make mistakes? The examples we have seen so far suggest otherwise. So perhaps we need a better term, one which also does a better job of warning users. "Confabulations" may be a much better term. Confabulations are surrounded by the truth and are created to fit in with them. The more we trust the speaker, the less likely we are to notice confabulations. The more everything around the confabulation is convincing, the easier it is for the confabulation to hide from us. Confabulations are much more dangerous, and perhaps this term better captures the dangers of LLMs.

NOTES

1. Sebastian Kobler, Matthew Clemson, Angela Sun, and Jonathan K. Kummerfeld, "Your Students Don't Use LLMs Like You Wish They Did," preprint, arXiv, April 26, 2026 (submitted to ACL 2026), https://doi.org/10.48550/arXiv.2604.23486.

2. James Hutson, "Rethinking Plagiarism in the Era of Generative AI," *Journal of Intelligent Communication* 3, no. 2 (2024): 20–31, https://doi.org/10.54963/jic.v3i2.220.

3. Chunpeng Zhai, Santoso Wibowo, and Lily D. Li, "The Effects of Over-Reliance on AI Dialogue Systems on Students' Cognitive Abilities: A Systematic Review," *Smart Learning Environments* 11 (2024): 28, https://doi.org/10.1186/s40561-024-00316-7.

4. Ivar Frisch and Mario Giulianelli, "LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models," preprint, arXiv, February 5, 2024, https://doi.org/10.48550/arXiv.2402.02896.

5. For example, Tom Duenas and Diana Ruiz, "The Risks of Human Overreliance on Large Language Models for Critical Thinking," preprint, ResearchGate, November 2024, https://doi.org/10.13140/RG.2.2.26002.06082; Aimé Anita, "Digital Literacy in the 21st Century: Preparing Students for the Future Workforce," *Research Output Journal of Arts and Management* 3, no. 3 (2024): 40–43, https://www.researchgate.net/publication/383557214_Digital_Literacy_in_the_21st_Century_Preparing_Students_for_the_Future_Workforce.

6. For the importance of metacognitive awareness and epistemic vigilance in AI use, see Samia Moustaghfir and Hind Brigui, "Navigating Critical Thinking in the Digital Era: An Informative Exploration," *International Journal of Linguistics, Literature and Translation* 7, no. 1 (2024): 137–43, https://doi.org/10.32996/ijllt.2024.7.1.11x.

7. For example, see Ying Ying Lim and Paul Röttger, "Bias in the East, Bias in the West: A Bilingual Analysis of LLM Political Bias on U.S.–China-Related Issues," in *Findings of the Association for Computational Linguistics: EACL 2026* (Association for Computational Linguistics, 2026), 2301–26, https://doi.org/10.18653/v1/2026.findings-eacl.122; and Ju-Chun Ko, "Bilingual Bias in Large Language Models: A Taiwan Sovereignty Benchmark Study," preprint, arXiv, February 6, 2026, https://arxiv.org/html/2602.06371v1.

8. Manisha Mondal, "What Happens in a World Run by AI? They Fall in Love, Kill Themselves, Commit Arson," *ThePrint*, May 15, 2026, https://theprint.in/feature/what-happens-in-a-world-run-by-ai-they-fall-in-love-kill-themselves-commit-arson/2932227/.

9. Dogus Yuksel, Mehmet Cem Catalbas, and Bora Oc, "Language-Dependent Political Bias in AI: A Study of ChatGPT and Gemini," preprint, arXiv, April 8, 2025, https://doi.org/10.48550/arXiv.2504.06436.

10. As is suggested by this study: Suyash Fulay, William Brannon, Shrestha Mohanty, et al., "On the Relationship Between Truth and Political Bias in Language Models," in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics, 2024), 9004–18, https://doi.org/10.18653/v1/2024.emnlp-main.508.

11. Edinburgh TV Festival, "The James MacTaggart Lecture 2022: Emily Maitlis," YouTube video, 56:42, posted August 24, 2022, https://www.youtube.com/watch?v=e0dRBtSAvLM.

12. This example is inspired by the work of Ryo Kamoi, Sarkar Snigdha Das, Renze Lou, et al., "Evaluating LLMs at Detecting Errors in LLM Responses" (paper presented at the Conference on Language Modeling [COLM], Philadelphia, PA, October 2024), https://arxiv.org/pdf/2404.03602.

13. Gunnar Ridderström, "Librarian Finds 'Preposterous Number' of Fake References in Paper from Springer Nature Journal," *Retraction Watch*, March 6, 2026, https://retractionwatch.com/2026/03/06/librarian-finds-preposterous-number-of-fake-references-in-paper-from-springer-nature-journal/.

14. Jiří Přibyl, Magdalena Krátká, and Michaela Tichá, "When AI Teaches Geometry Wrong: Systematic Errors in LLM-Generated Explanations and Their Educational Risks," in *ICERI2025 Proceedings* (IATED, 2025), 4760–69, https://doi.org/10.21125/iceri.2025.1357.

15. Mrinank Sharma, Meg Tong, Tomasz Korbak, et al., "Towards Understanding Sycophancy in Language Models," preprint, arXiv, last revised May 10, 2025, https://doi.org/10.48550/arXiv.2310.13548.

16. Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky, "ELEPHANT: Measuring and Understanding Social Sycophancy in LLMs," preprint, arXiv, last revised September 29, 2025, https://doi.org/10.48550/arXiv.2505.13995.

17. Shan Chen, Mingye Gao, Kuleen Sasse, et al., "When Helpfulness Backfires: LLMs and the Risk of False Medical Information Due to Sycophantic Behavior," *npj Digital Medicine* 8 (2025): 605, https://doi.org/10.1038/s41746-025-02008-z.

18. Harry G. Frankfurt, *On Bullshit* (Princeton University Press, 2005).

19. Michael Townsen Hicks, James Humphries, and Joe Slater, "ChatGPT Is Bullshit," *Ethics and Information Technology* 26 (2024): 38, https://doi.org/10.1007/s10676-024-09775-5.

20. James Humphries, Michael Townsen Hicks, and Joe Slater, "LLMs Bullshit by Design: A Reply to Licon," *Philosophy & Technology* 39, no. 2 (2026): 98, https://doi.org/10.1007/s13347-025-01016-x.

21. Joe Slater, James Humphries, and Michael Townsen Hicks, "AI's Bullshitting Obscures Who's to Blame for Its Mistakes," *Scientific American*, July 17, 2024, https://www.scientificamerican.com/article/chatgpt-isnt-hallucinating-its-bullshitting/.

A Contemporary Footnote to Plato: Academic Education in the Age of AI-Related Cognitive Offloading

Isaac Nevo
BEN-GURION UNIVERSITY OF THE NEGEV

> But when he came to letters, This, said Theuth, will make the Egyptians wiser and give them better memories; it is specific both for the memories and for the wit. Thamus replied: O most ingenious Theuth, the parent or inventor of an art is not always the best judge of the utility or inutility of his own inventions to the users of them. And in this instance, you who are the father of letters [*grammata*], from a paternal love of your own

> children have been led to attribute to them a quality which they cannot have; for this discovery of yours will create forgetfulness in the learners' souls, because they will not use their memories; they will trust the external written characters and not remember themselves.[1]

It is by now almost a commonplace that AI tools that employ large language models (LLMs) to simulate human intelligence in various products create not only positive opportunities for their users but, as well, serious challenges to those users. Among the challenges, the promise of greater "intelligence" in various AI products may come at the expense, for the users of such tools, of both *lesser* intelligence and *lowered* cognitive capacity.

The problem has come to be described as one of "cognitive offloading," which is the result of outsourcing cognitive tasks ("loads") by means of a readily available technology that diminishes the user's autonomous skills in performing the cognitive task in question. Cognitive offloading results—or it often leads to—cognitive erosion, specifically, a decline in the user's cognitive abilities (and consequently in her autonomy). Thus, for example, Michael Gerlich describes, on the basis of his research findings, "a significant negative correlation between the frequent use of AI tools and critical thinking abilities, mediated by the phenomenon of cognitive offloading."[2] The expression "critical thinking abilities" in this context signifies the autonomous capacity to evaluate the truth or falsity of statements, the validity or invalidity of inferences, and the justification or lack thereof in other forms of reasoning. In the context of education, both generally, and specifically in academia, this negative aspect of the multiple uses of AI should raise serious educational and ethical concerns. This is not to deny that cognitive offloading may also have positive outcomes (such as freeing mental space, or brain activity, for other tasks or challenges). But the negative impact is still there, especially when the offloading is persistent, and the cognitive tasks that are offloaded are of the more important kind.[3]

In this paper, I consider the problem of cognitive offloading as it pertains to the educational role of universities and colleges. If encouraging the use of AI tools by students in higher education leads to their cognitive impoverishment, and consequently to diminishing their skills and expertise, then, clearly, we, as administrators or faculty of such institutions, face a dilemma. The promise of education, academic or otherwise, is the cognitive *empowerment* of students, i.e., making them more capable of thinking, writing, computing, researching, etc. *on their own*, both for its own sake and as a power that will serve the students well in their subsequent lives and careers, and also for the sake of society at large, where the students may come to serve as experts and authorities in their respective fields. Cognitive erosion due to technological offloading could render this "promise," which is a central aim of higher education, hollow. Modern societies have come to rely on institutions of higher education for certifying the credentials of the experts that these societies come to rely on, but this role of higher education and the actual value of the academic degrees it confers stands to be undermined. Clearly, educational and academic institutions cannot ignore the development of AI, as it increasingly becomes an inevitable part of the training required in contemporary society, for its promise of gains in access to data, in computing power, and in various forms of research. However, the burden of cognitive offloading, specifically, the decline in the students' autonomous cognitive capacities, will be borne not only by the students themselves but, as well, by the rest of society in *its* reliance on academic experts.

Philosophers will recall that the problem is hardly new. In the *Phaedrus*, Plato introduces the myth of Theuth, an Egyptian god who was "the inventor of many arts," including "arithmetic and calculation and geometry and astronomy and draughts and dice," but whose "great discovery was the use of letters," i.e., writing and reading: literacy. Seeking permission to disseminate these discoveries to the people of Egypt, Theuth presented them to the god king of Egypt Thamus (also known as Ammon) for evaluation. In Plato's telling, Thamus responded critically to the invention, pointing to a potential drawback of the widespread use of writing. Contrary to Theuth's claim that the use of writing would enhance the users' memory and wit—which are cognitive powers on which wisdom and learning depend—Thamus argued that the use of writing will "create forgetfulness in the learners' souls," as these learners will come to rely on external devices in their mnemonic tasks, while neglecting their own, independent, cognitive capacity of remembering (and thinking). As a (technological) aid to cognition, the art of writing involves the disadvantage (or risk) of cognitively "de-skilling" its users, that is, of eroding their *own* capacities, thus leaving them dependent on external devices at the expense of their own cognitive autonomy.

Three points are worth noting. First, some negative impacts are often associated with new technologies or "arts," and a balance should be struck between "utility or inutility," i.e., the advantages and disadvantages of the proposed change. This is obviously true regarding *any* technology. However, one should keep in mind that the inventors of proposed new technology are not themselves best suited to notice its drawbacks. Secondly, in evaluating a technology, particularly its advantages, the inventor of the technology is prone to "attribute to them a quality which they cannot have." In particular, the tendency to project expected positive outcomes of the proposed new technology as inherent qualities of the technology in question is a common mistake that could serve to obscure the disadvantages of the use of that technology. Additionally, external aids to internal (cognitive) functioning may induce a problem internally, e.g., a "forgetfulness in the learner's souls," or in our terms, cognitive decline due to technologically induced cognitive outsourcing. As we shall see, these points are relevant for the evaluation of technologies recently presented, particularly AI, the negative impacts of which could result in cognitive erosion—a negative effect that may be obscured by the tendency to attribute intelligence to what is actually and merely an artificial device.

Of course, judgments may differ with respect to each of Plato's points regarding different specific technologies. It may be disputed whether dependence on a particular

technology leads to any decline in the users' own skills and capacities. Thus, for example, it may be questioned whether GPS-based apps such as Waze erode their users' autonomous navigation skills while successfully directing them through the urban maze. Furthermore, even if that *is* found to be the case, many users of Waze might welcome this cognitive trade for the benefit of easier and more economical travel.

A more general Platonic complaint against the introduction of writing lurks in the background and motivates Plato's invocation of the myth of Theuth. It is Plato's complaint about rhetoric, as practiced by the sophists of his day (which is the main theme of the *Phaedrus)*. In rhetorical compositions of the kind that Plato's character in the *Phaedrus* admires, words are combined, and presented to their audiences, regardless of their meanings, or true definitions, forming pleasant speeches that have little or no relation to the truth of what is said. As a result—and as Socrates never tired of pointing out—such compositions not only produced no new knowledge, but the people delivering them, Plato claimed, had absolutely no idea of what they were saying. Adapting Thamus's language, rhetoric produces "unthinking in the users' souls." So, here it is not only memory that is lost by offloading to "external" devices, but the very capacity to think in terms of truth and falsity that is lost by "offloading" thinking to rhetoric. (As we shall see, a similar point can be made regarding the use of AI.)

In recent discussion of AI use and the potential damages from its use, there has been a tendency to view Plato's story as a refutation—in the form of a *reductio-ad-absurdum*—of the worries expressed above regarding some cognitive decline due to offloading. After all, so the argument goes, there was hardly any *greater* cognitive *enhancement* than that introduced by the alphabet, namely, the power of literacy. So the worry about "forgetfulness in the soul" should be dismissed. The advantages of the "technology" of reading and writing far outweigh the disadvantages (if any) to memory as a cognitive capacity. What is lost in the erosion of one skill or capacity is offset by the development of other, technologically driven, skills and capacities. But while Plato has certainly not done justice to the positive impact of literacy, his point is well worth considering regarding other technologies and *their* power to erode cognition. The "utility or inutility" of various inventions in terms of enhancing or diminishing cognition is a theme that should be considered separately in each case. *Some* technologies may offer greater advantages than disadvantages in terms of cognitive enhancement, but others may not: the calculus of utility may be different in different cases. In particular, the use of generative AI may not offer as much as literacy does in the way of cognitive enhancement, though it certainly does offer surprising new tools and benefits, and in doing this it may also threaten a level of cognitive erosion well beyond "forgetfulness in the soul," one that could leave its users without crucial capacities for intellectual functioning. By inducing "unthinking in the users' souls," the use of artificial intelligence may leave us a lot less intelligent. This is the dilemma—let's call it "Plato's dilemma"—for higher education when it comes to the ever more pervasive use of AI tools in the processes of higher learning.

We should note that the problem of cognitive offloading is not the only ethical problem that the use of AI creates in the sphere of education. Large language models (LLMs) are trained on written materials that are available on the internet without the permission of its authors (or right holders), and so without due credit to their authors. This puts the users of its products at risk of "secondary" plagiarism, i.e., of using the material of others without permission of and/or credit to those others. Thus, the dependence on AI on the part of students and/or faculty may blur the distinction between work that is one's own and work rightly attributed to others, thus encouraging forms of cheating on student assignments, and perhaps even duplicity on the part of faculty with respect to work presented for publication. Clearly, proper documentation is required to avoid false impressions, but, just as clearly, opportunities for dishonest work have been greatly enhanced by the availability of AI. As a consequence, faculty may retreat from assessing students on the basis of written work that they have *not* done in class (e.g., seminar or term papers done as *homework*) and move to assess students only through work that the students have done under the watchful eye of the teacher—for example, in-class written essays and/or exams). Unfortunately, doing so risks doing away with one of the primary methods of academic teaching, namely, library (and/or archive) research. For this reason, this mode of teaching—requiring for the academic assessment of students that they be judged solely on what they produce in the classroom—has been met with the criticism that such a policy renders more harm than good for it "throws the baby out with the bathwater" as students will no longer be trained to do their own independent research and to formulate their own ideas in the process of their research. Clearly, these are legal and ethical issues that will have to be resolved, in years to come, by academic institutions, by journal editors, and by other stakeholders in society.

Plato's second point regarding the error of attributing cognitive powers to technological aids, which, he rightly claims, is "a quality which they cannot have" is powerfully exemplified in current debates over AI. Artificial *intelligence* is judged to be "intelligent" because its products *simulate* intelligence when, as a matter of fact, these apparently intelligent products are produced by mechanisms that exhibit neither independent judgment of truth or falsity nor any reasoning regarding the subject matter at hand. Only statistical manipulations of word sequences are responsible for the outcome, and while the products may be very impressive, they are not the products of genuine intelligence. As noted, Plato's complaint against rhetoric as practiced by various sophists is that by rhetoric words are combined without regard to their precise meanings, or definitions, but rather by their non-referential features, such as their sound and their aesthetic properties (as one might find, say, in poetry or in a speech intended to have emotional appeal). With artificial intelligence we have a similar phenomenon: words are combined not based on their meanings or references but by their statistical properties, that is, the likelihood of their occurring next to other words. The outcome may be very impressive, but it is not the fruit of genuine thinking on the subject matter at hand, as testified by the prevalence of (what is referred to as) AI "hallucinations" and the unthinking

proliferation of biases. Nevertheless, by calling artificial intelligence "intelligent," the crucial distinction between *true* intelligence and *mechanical reproduction* is obscured.

The following is an argument for the claim above. A minimal requirement for the possession of cognitive intelligence is the ability to judge, correctly or mistakenly, the truth or falsity of a proposition. That will be our first assumption. (A minimal requirement of *practical* intelligence is the ability to adapt means to ends, which is found in some animals who clearly lack cognitive intelligence. In the following I shall focus on cognitive intelligence.) Thus,

1. For all x, if x is cognitively intelligent, then x can judge, correctly or not, the truth or falsity of a proposition. (A premise)

2. If Chat (an AI tool) is cognitively intelligent, then Chat can judge, correctly or not, the truth or falsity of a proposition. (by Universal Instantiation on # 1)

To be sure, judgments of truth or falsity on the part of the cognitively intelligent are not immune to mistakes of various kinds. But the possibility of a mistake with respect to a particular claim presupposes the capacity, on the part of the person making the claim, to distinguish between truth and falsity and to aim at the former. Indeed, the ability to judge the truth or falsity of a proposition is a complex ability involving the ability to understand the proposition in question, i.e., the sense and reference of the terms used, their combination into the proposition in question, and the truth value of the claim that is being made. Clearly, a system operating on words only, creating sequences on statistical bases without regard to meanings or references (that is, without having a representation of the world) does not have the required set of capacities. Hence: We may assume that:

3. Chat (an AI tool) cannot judge, correctly or not, the truth or falsity of a proposition. (A premise)

Therefore:

4. Chat is not cognitively intelligent. (by Modus Tollens on # 3 and # 2)

Of course, our friend Chat can do some amazing things. It can produce answers to many questions we put to it, and respond to many prompts that we give it, with answers that are often correct and enlightening. It thereby simulates judgments of truth and falsity without genuinely making them. It can uncover hidden knowledge and give us access to a multitude of data we would not otherwise have access to. Nevertheless, it does so without intelligence, and without appreciation of either the meanings of the word sequences it produces or the truth or falsity of the sentences it combines. Like Plato's sophists, it simulates cognitive intelligence rather than exhibits it. (Admittedly, this is a bit unfair to the sophists who did not operate *exclusively* on purely rhetorical principles and so exhibited cognitive and practical intelligence in addition to their rhetorical skills.)

The premises of the above argument may, of course, be disputed. Premise 1 is a restatement of the claim that being able to judge truth or falsity is a requirement of intelligence. But perhaps being able to judge the truth or falsity of a proposition is not a necessary condition of intelligence and intelligence can be exhibited in lesser undertakings, e.g., the ability to manipulate words, generate sentences, and compute the probabilities of certain sequences. These capacities, which are clearly exhibited by AI tools, are mechanizable. But who is to say that cognitive intelligence *isn't*? After all, in humans, too, intelligence rests on neural mechanisms of one sort or another. No ghost is secretly operating the machine with us humans. Nevertheless, the idea of (cognitive) intelligence without genuine cognition, i.e., without moving from word to meaning, from sentence to proposition, and from proposition to world, leaves something to be desired. It leaves the cognitive aspect of intelligence behind, and that is a stiff price to pay. Cognitive intelligence without cognition—that is, without reasoning—appears to be an oxymoron.

A similar point was made by Shannon Vallor with reference to the philosopher Willfred Sellars's concept of "the space of reason":

> In 1956, Wilfrid Sellars spoke of knowledge as TAKING PLACE in "a logical space of reasons, of justifying and being able to justify what one says." This means that to be knowers—and more broadly to be intelligent in the ways humans characteristically are—our minds must be more than inert repositories of facts or true statements. Our minds must be active and competent movers within the space of reasons.[4]

And, referring to the paper by Heinrichs and Knell, "Aliens in the Space of Reasons?," she goes on to say:

> The kinds of computational tools that we call AI are wholly incapable of standing with us in the space of reasons. One of the most powerful yet dangerous aspects of complex machine learning models is that they can derive solutions to knowledge tasks in a manner that entirely bypasses this space.[5]

Premise 3 above (namely, that "Chat [an AI tool] cannot judge, correctly or not, the truth or falsity of a proposition") could also be challenged. To be sure, current AI tools operate, roughly, by calculating statistical probabilities among word sequences, and that may be insufficient for judgments of truth or falsity. Still, a future AGI (artificial general intelligence) tool, operating on some *other* principle, might close that gap. Such a tool may one day be both wholly mechanical *and* capable of assessing truth or falsity. Perhaps so, but for now this remains speculative. At present, the prevalence of AI "hallucinations" and biases suggests a lack of capacity to distinguish truth from falsity.

Coming back to our main theme, we can now see that by encouraging our students to use generative AI tools, and to rely on these tools for research and writing, we may be encouraging them to offload cognitive tasks by

means of a technology that merely simulates intelligence. Unfortunately, projecting intelligence upon such devices may serve to obscure the unthinking nature of the process by which its products have been produced, as well as the resulting unthinking in the user's mind, just as projecting the power of memory enhancement upon Theuth's letters could serve to obscure the resulting forgetfulness in the soul. The true balance of "utility or inutility" tends to become hidden from sight.

The character of the technology that we use enters the calculation of "utility and inutility" of cognitive offloading in other ways as well. As noted, Plato's general attack on the Sophists points to a risk regarding rhetorical uses of written language. Rhetoric creates opportunities for combining words without reference to their true definitions. Nevertheless, writing creates opportunities for cognitive enhancement far outweighing the risk of rhetoric or forgetfulness. By contrast, the benefits of AI tools are mostly in the products these tools create, and not in the users' cognition. AI tools regularly combine words statistically, without regard to their meaning and reference, risking not just forgetfulness, but much broader unthinking, or cognitive decline, in their users' minds. Letters are not themselves quasi-sophists; chatbots are **precisely** that. Plato's judgment that the art of writing (as well as literacy in general) may induce "forgetfulness in the soul" may be countered by pointing to the great utility of literacy in terms of enhancing not only memory but also learning, thinking, scholarship, science, and general enlightenment. One has only to think of the many cognitive skills that we now have but were not there before we "off loaded" memory to written documents: the capacity to read and write and thereby to interpret, think about, and discuss with others the documents we produce. By contrast, cognitive offloading by means of AI technology makes us dependent on an intelligence-simulating machine that replaces the need to think independently in terms of truth and falsity. Unlike old Theuth's letters, we offload by means of a technology that erodes not only memory but intelligence itself, specifically, the disposition to think of what we hear and read in terms of the truth and falsity of what we have heard and/or read.

So how are we to deal with Plato's dilemma in contemporary academe? Given that there is no question of our avoiding use of AI altogether (or luddite-like smashing the AI machines), how are we to avoid the creation of unthinking in our students' souls? While there is no single answer, and no single approach that would fit all, here are some suggestions.

First, the problem of cognitive offloading is independent of other ethical concerns. It occurs even where there is no actual cheating or dishonesty on the part of students. It is an ethical issue for *us*, the faculty, for it is *we* who incur the risk of unwittingly cheating our students out of a genuine cognitive development. And as in the cases of academic dishonesty—with which we can only deal by assuming (perhaps naively) that most students do not want to cheat and will resist temptations to do so in a well-functioning academic environment—so with regard to offloading, it can (or should) be assumed that most students do care about acquiring skills and competences, and would feel cheated to learn that the education they paid for had actually *de*-skilled them even while granting them a degree. So, the first suggestion is to get the students involved in addressing the problem. This could be done by education's traditional methods: exposing them to the problem, assigning relevant reading materials to the students, debating with them about the appropriate use of AI, and warning them of the negative consequences of its unthinking use.

Secondly, and less collaboratively, methods of testing should be developed, and facilitated, that separate students from machine performance. This could be done by reverting to old tactics such as the "in class, blue-book exam," or the oral exam, where the point is to test competence directly, separating it from unwanted contributions. As noted above, this tactic risks "throwing the baby out with the bath water" as autonomous research and writing would no longer be part of the students' training. But perhaps this could be overcome by creating longer in-class "writing sessions" where students would be trained and tested on writing pieces longer than answers to in-class exam questions and do so on topics or questions of their own.

Thirdly, methods of instruction should be changed. One might, for example, ask students to do an assignment twice—first without the aid of AI and then with some level of reliance on an AI tool. Once the student has finished doing the assignment with the use of AI, the student should then compare and evaluate the difference. (Clearly, the AI product will be better in some respects, but the student will not be "stupefied" by the process. The better product could spur the students to do better themselves the next time.)

These suggestions are not offered as proven methods, and perhaps they, too, fall short of addressing the real challenge of cognitive offloading. The principles, however, should be further developed, both the principle of student involvement and the principle of separating competence from performance in both teaching and assessing. Plato's Socrates offered the dialogical method of the *elenchus*—*separating* cognitive competence from linguistic performance—as a remedy to sophistry. That, too, where applicable, could be useful.

NOTES

1. Plato, *Phaedrus* 274–75. All quotations from the *Phaedrus* in this article are from this passage.

2. Gerlich, "AI Tools in Society." For more on the theory of cognitive offloading see Risko and Gilbert, "Cognitive Offloading."

3. See also Kosmyna et al., "Your Brain on ChatGPT," for a study of the effects of cognitive offloading at the level of brain activity. Their study looked at three groups performing a writing task, one working without technological aid, one using an AI tool, and one using a search engine. Electroencephalography (EEG) was used to record participants' brain activity in each of these groups in order to assess their cognitive engagement and cognitive load. The study found that brain activity was lowest in the AI-using group. The authors conclude: "In this study we demonstrate the pressing matter of a likely decrease in learning skills based on the results of our study. The use of LLM had a measurable impact on participants, and while the benefits were initially apparent, as we demonstrated over the course of 4 months, the LLM group's participants performed worse than their counterparts in the Brain-only group at all levels: neural, linguistic, scoring" (2).

4. Vallor, *AI Mirror*, 106–7. Vallor is referring to Sellars, "Empiricism and the Philosophy of Mind."

5. Vallor, *AI Mirror*, 107. Vallor is referring to Heinrichs and Knell, "Aliens in the Space of Reasons?"

BIBLIOGRAPHY

Gerlich, Michael. "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking." *Societies* 15, no. 1 (2025): 6. https://doi.org/10.3390/soc15010006.

Heinrichs, Bert, and Sebastian Knell. "Aliens in the Space of Reasons? On the Interaction Between Humans and Artificial Intelligent Agents." *Philosophy & Technology* 34 (2021): 1569–80. https://doi.org/10.1007/s13347-021-00475-2.

Kosmyna, Nataliya, Eugene Hauptmann, Ye Tong Yuan, et al. "Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task." Preprint, arXiv, last revised December 31, 2025. https://doi.org/10.48550/arXiv.2506.08872.

Plato. *Phaedrus*. In *The Dialogues of Plato*, translated by Benjamin Jowett. Great Books of the Western World 7. Encyclopedia Britannica, 1952.

Risko, Evan F., and Sam J. Gilbert. "Cognitive Offloading." *Trends in Cognitive Sciences* 20, no. 9 (2016): 676–88. https://doi.org/10.1016/j.tics.2016.07.002.

Sellars, Wilfrid. "Empiricism and the Philosophy of Mind." In The Foundations of Science and the Concepts of Psychology and Psychoanalysis, edited by Herbert Feigl and Michael Scriven, 253–329. Minnesota Studies in the Philosophy of Science 1. University of Minnesota Press, 1956.

Vallor, Shannon. *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. Oxford University Press, 2024. https://doi.org/10.1093/oso/9780197759066.001.0001.