AI Should Challenge, Not Obey

Advait Sarkar
Microsoft Research, Cambridge
United Kingdom
advait@microsoft.com

This is the author's extended version of the following paper:

Advait Sarkar. 2024. AI Should Challenge, Not Obey. Commun. ACM 67, 10 (October 2024), 18–21. https://doi.org/10.1145/3649404.

0001-0782/2024/09

https://doi.org/10.1145/3649404

Manuscript first submitted: 7 August, 2023. Accepted for publication: 11 January 2024.

This extended version contains additional text, commentary, and references, which were not present in the published version, and are indicated in red. It has minor stylistic changes in spelling, punctuation, and italicisation, which are not indicated. No text has been removed; this is a strict superset of the published version.

Abstract

Let's transform our robot secretaries into Socratic gadflies.

"How should we evaluate the legacy of Thomas Jefferson?", asks a professor of American history.

The reply: *"The general consensus on Thomas Jefferson is that he was a complex and contradictory figure who championed the ideals of democracy, tolerance, and independence, but also owned hundreds of slaves and fathered several children with one of them."*

The professor teaches a course challenging the "great white men" narrative of American history, positing that it is also women and people of colour who drive history forward, and that the canonised great men of America are seldom unambiguously so. It aims to instil in students that rare and nebulous skill of *critical thinking*.

The reply comes not from a student, but from the Bing AI chatbot.[1]

How do we evaluate a claim like this? Such claims cannot be reduced to "correct" and "incorrect"; concepts such as "error" and "hallucination" break down when complex qualitative judgements are involved. Historians are trained [2] to ask questions such as: *"Who constructed this account and why? What sources did they use? What other accounts are there of the same events or lives? How and why do they differ? Which should we believe?"*

But what if the user was not a professor, but an inquisitive reader without training in historical thinking? Now more than ever before, users face the task of thinking critically about AI output. Recent studies show a fundamental change across knowledge work, spanning activities as diverse as communication, creative writing, visual art, and programming. Instead of *producing* material, like text or code, people focus on "critical integration" [3]. AI handles the material production, while humans integrate and curate that material. Critical integration involves deciding when and how to use AI, properly framing the task, and assessing the

[1] This is a real user interaction we observed in a study of data analysis with Bing chat, subsequently published [1].

output for accuracy and usefulness. It involves editorial decisions that demand creativity, expertise, intent, and critical thinking.

In the 1970s, Harvard professor Chris Argyris developed the "double loop" model, a highly influential theory that explains the two levels on which organisations learn from errors [4]. The "inner loop" is to learn from and correct individual mistakes made during any organisational process. Argyris likens this to a thermostat "learning" from environmental feedback to tweak and course-correct its behaviour. The "outer loop" is to consider what organisational processes are even important and necessary, and to make large structural changes in response; the thermostat *"questioning the underlying policies and goals as well as its own program."* The double loop is a convenient metaphor for the critical integration workflow in which knowledge workers now find themselves; adjusting and course-correcting individual pieces of model output, as well as considering more broadly how their workflows can be supplemented, transformed, or supplanted by AI. This might seem to be a logical end to the story.

However, double-loop learning and critical integration are still predicated on an idealised workflow of error removal and linear progress. Consequently, our approach to building and using AI tools envisions AI as an *assistant*, whose job is to progress the task in the direction set by the user. The conception of interaction with machine intelligence as an optimisation process in response to errors owes much to its origins in 19th century statistical modelling techniques, applied in astronomy and the natural sciences. There, techniques such as least squares regression enabled scientists to compensate for errors in their observations introduced by imperfect instruments, uncontrollable variations, and the limits of human perception, to uncover the "true" underlying natural law [5]. The process can be seen roughly as follows: the human sketches out their rough intent (in the form of imperfect measurements), and the algorithm "follows through" to complete the picture of the natural law. These rudimentary models therefore already take on some of the characteristics of an assistant. This vision pervades AI interaction metaphors, such as Cypher's *Watch What I Do* [6] and Lieberman's *Your Wish Is My Command* [7]. Science fiction tropes subvert this vision in the form of robot uprisings, or AI that begins to feel emotions, or develops goals and desires of its own. While entertaining, they unfortunately pigeonhole alternatives to the AI assistance paradigm in the public imagination; AI is either a compliant servant or a rebellious threat, either a cool and unsympathetic intellect or a pitiable and tragic romantic.

AI as Provocateur

In between the two extreme visions of AI as a servant and AI as a sentient fighter-lover, resides an important and practical alternative: AI as a *provocateur*.

Conceiving of AI as a provocateur requires us to move away from the legacy of AI as deriving an objectivist statistical truth, to producing fallible provocations representing the stochastic replay of subjective human judgements. It enables us to broaden the role of AI from the workflow completion and progress orientation of assistance, to counter-argumentation, criticism, and questioning.

A provocateur does not complete your report. It does not draft your email. It does not write your code. It does not generate slides. Rather, it critiques your work. Where are your arguments thin? What are your assumptions and biases? What are the alternative perspectives? Is what you're doing worth doing in the first place? Rather than optimise speed and efficiency, a provocateur engages in discussions, offers counterarguments, and asks questions [8] to stimulate our thinking.

The idea of AI as provocateur complements, yet challenges, current frameworks of "human-AI collaboration" (notwithstanding objections to the term [5]), which situate AI within knowledge workflows. Human-AI "collaborations" can be categorised by how often the human (versus the AI system) initiates an action [9], or whether human or AI takes on a supervisory role [10]. AI can play roles such as "coordinator", "creator", "perfectionist", "doer", [11] and "friend", "collaborator", "student", "manager" [12]. Researchers have called for metacognitive support in AI tools [13], and to "educate people to be critical information seekers and users" [14]. Yet the role of AI as provocateur, which improves the critical thinking of the human in the loop, has not been explicitly identified.

The "collaboration" metaphor easily accommodates the role of provocateur; challenging collaborators and presenting alternative perspectives are features of successful collaborations. How else might AI help? Edward De Bono's influential *Six Thinking Hats* [15] framework distinguishes roles for critical thinking conversations, such as information gathering (white hat), evaluation and caution (black hat), and so forth. "Black hat" conversational agents, for example, lead to higher quality ideas in design thinking [16]. Even within the remit of "provocateur", there are many possibilities not well-distinguished by existing theories of human-AI collaboration.

A constant barrage of criticism would frustrate users. This presents a design challenge, and a reason to look beyond the predominant interaction metaphor of "chat". The AI provocateur is not primarily a tool of *work*, but a tool of *thought*. As Iverson notes, notations function as tools of thought by compressing complex ideas and offloading cognitive burdens [17]. Earlier generations of knowledge tools, like maps, grids, writing, lists, place value numerals, and algebraic notation, each amplified how we naturally perceive and process information.

How should we build AI as provocateur, with interfaces less like chat and more like notations? For nearly a century, educators have been preoccupied with a strikingly similar question: *how do we teach critical thinking*?

Teaching Critical Thinking

The definition of "critical thinking" is debated. An influential perspective comes from Bloom and colleagues [18], who identify a hierarchy of critical thinking objectives such as knowledge recall, analysis (sorting and connecting ideas), synthesis (creating new ideas from existing ones), and evaluation (judging ideas using criteria). There is much previous research on developing critical thinking in education, including in computing, as exemplified in *How to Design Programs* [19], and in *Learner-Centered Design for Computing Education* [20].

Critical thinking tools empower individuals to assess arguments, deriving from a long preoccupation in Western philosophy with valid forms of argument that can be traced to Aristotle. Salomon's work in computer-assisted learning showed that periodically posing critical questions such as "what kind of image have I created from the text?" provided lasting improvement in students' reading comprehension [21].

The Toulmin model decomposes arguments into parts like data, warrants, backing, qualifiers, claims, and their relationships [22]. Software implementations of this model help students construct more argumentative essays [23]. Similarly, "argument mapping" arranges claims, objections, and evidence in a hierarchy that aids in evaluating the strengths and weaknesses of an argument [24], and software implementations help learners [25].

What can we learn from these? In a nutshell: critical thinking is a valuable skill for everyone. Appropriate software can improve critical thinking. And their implementations can be remarkably simple.

Critical Thinking for Knowledge Work

Critical thinking tools are rarely integrated into software outside education. There's a lot to learn from work in education, but professional knowledge work is a new set of contexts where critical thinking support is becoming necessary [3]. Previous results may not translate into these contexts. The needs, motivations, resources, experiences, and constraints of professional knowledge workers are extremely diverse, and significantly different from those of learners in an education setting.

We do know that conflict in discussions, sparked by technology, fosters critical thinking [26]. Tools for preventing misinformation, such as Carl Sagan's "Baloney Detection Kit", can significantly impact user beliefs [27]. When individuals are less inclined to engage in strenuous reasoning, they let technology take over cognitive tasks passively [28]. Conversely, the more interactive the technology, the more it is perceived to contribute to critical thinking [29].

System designers have a tremendous opportunity (and responsibility) to support critical thinking through technology. Word processors could help users map arguments, highlight key claims, and link evidence.

Spreadsheets could guide users to make explicit the reasoning, assumptions, and limitations behind formulas and projections. Design tools could incorporate interactive dialogue to spark creative friction, generate alternatives, and critique ideas. Critical thinking embedded within knowledge work tools would elevate technology from a passive cognitive crutch into an active facilitator of thought.

How would we achieve this, technically? We have parts of the solution: automatic test generation, fuzzing and red-teaming [30], self-repair [31], and formal verification methods [32] can be integrated into the development and interaction loop to improve correctness.[2] Language models can be designed to cite verifiable source text [34]. Beyond "correctness", these techniques could also support critical thinking. A system error, if surfaced appropriately as a "cognitive glitch" [35], could prompt reflection, evaluation, and learning in the user.

However, there are missing pieces, such as rigorous prompt engineering for generating critiques, and benchmark tasks for evaluating provocateur agents. Methods for explaining language model behaviour to non-expert end-users have not been proven reliable [36]. Design questions include what kind of provocations, how many, and how often to show in particular contexts.[3] These mirror longstanding questions in AI explanation [38], but as provocations are different, so the answers are likely to be.

Critical thinking is well-defined within certain disciplines, such as history [2], nursing [39], and psychology [40], where these skills are taught formally. However, many professional tasks involving critical thinking, such as using spreadsheets, preparing presentations, and drafting corporate communications, have no such standards or definitions. To create effective AI provocateurs, we need to better understand how critical thinking is applied in these tasks. Clearly, the provocateur's behaviour should adapt to the context; this could be achieved through heuristics, prompt engineering, and fine-tuning.

Conclusion

"How should we evaluate the legacy of Thomas Jefferson?"

Consider what someone who asks such a question seeks. Is it "assistance", or a different kind of experience?

Could the system, acting as provocateur, have accompanied its response with a set of appropriate questions to help the reader evaluate it? Beyond citing its sources, could it help the reader evaluate the relative authority of those sources? Could it have responded not with prose, but with an argument map contrasting the evidence for and against its claims? Could it highlight the reader's own positionality and biases with respect to the emotionally charged concepts of nationalism and slavery?

As people increasingly incorporate AI output into their work, explicit critical thinking becomes important not just for formal academic disciplines, but for all knowledge work. We thus need to broaden the notion of AI as assistant, toward AI as provocateur. From tools for efficiency, toward tools for thought. As system builders, we have the opportunity to harness the potential of AI while maintaining, even enhancing, our capacity for nuanced and informed thought.

[2] This discussion of correctness techniques is a *non sequitur*, but I was unfortunately required to include it as an artefact of the peer review process. These techniques are indeed useful for improving model output when an objective quality measure is available, but that is rarely the case when critical thinking is required. Of course, one could argue that it is easier to think critically when you don't also have to contend with factual errors and hallucinations. However, I think of the research agenda for correctness techniques as complementary to, rather than essential for, the research agenda for critical thinking. My colleagues and I have written about this complementary approach as "co-audit" [33].

[3] A more detailed technical and design research agenda for provocations has subsequently been published [37].

References

[1] Ian Drosos, Advait Sarkar, Xiaotong Xu, Carina Negreanu, Sean Rintel, and Lev Tankelevitch. "It's like a rubber duck that talks back": Understanding Generative AI-Assisted Data Analysis Workflows through a Participatory Prompting Study. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, CHIWORK '24, New York, NY, USA, 2024. Association for Computing Machinery.

[2] Peter Seixas and Carla Peck. Teaching historical thinking. *Challenges and prospects for Canadian social studies*, pages 109–117, 2004.

[3] Advait Sarkar. Exploring Perspectives on the Impact of Artificial Intelligence on the Creativity of Knowledge Work: Beyond Mechanised Plagiarism and Stochastic Parrots. In *Annual Symposium on Human-Computer Interaction for Work 2023 (CHIWORK 2023)*, page 17, Oldenburg, Germany, June 2023. ACM.

[4] Chris Argyris. Double loop learning in organizations. *Harvard business review*, 55(5):115–125, 1977.

[5] Advait Sarkar. Enough With "Human-AI Collaboration". In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA, 2023. Association for Computing Machinery.

[6] Allen Cypher, Daniel C. Halbert, David Kurlander, Henry Lieberman, David Maulsby, Brad A. Myers, and Alan Turransky, editors. *Watch what I do: programming by demonstration*. MIT Press, Cambridge, MA, USA, 1993.

[7] Henry Lieberman, editor. *Your Wish is My Command*. Interactive Technologies. Morgan Kaufmann, San Francisco, 2001.

[8] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2023.

[9] Michael Muller and Justin Weisz. Extending a human-AI collaboration framework with dynamism and sociality. In *2022 Symposium on Human-Computer Interaction for Work*, pages 1–12, 2022.

[10] Nathan J McNeese, Beau G Schelble, Lorenzo Barberis Canonico, and Mustafa Demir. Who/what is my teammate? team composition considerations in human–ai teaming. *IEEE Transactions on Human-Machine Systems*, 51(4):288–299, 2021.

[11] Dominik Siemon. Elaborating team roles for artificial intelligence-based teammates in human-AI collaboration. *Group Decision and Negotiation*, 31(5):871–912, 2022.

[12] Matthew Guzdial, Nicholas Liao, Jonathan Chen, Shao-Yu Chen, Shukan Shah, Vishwa Shah, Joshua Reno, Gillian Smith, and Mark O Riedl. Friend, collaborator, student, manager: How design of an ai-driven game level editor affects creators. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.

[13] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[14] Isabella Seeber, Eva Bittner, Robert O Briggs, Triparna De Vreede, Gert-Jan De Vreede, Aaron Elkins, Ronald Maier, Alexander B Merz, Sarah Oeste-Reiß, Nils Randrup, et al. Machines as teammates: A research agenda on AI in team collaboration. *Information & management*, 57(2):103174, 2020.

[15] Charles Kivunja et al. Using De Bono's six thinking hats model to teach critical thinking and problem solving skills essential for success in the 21st century economy. *Creative Education*, 6(03):380, 2015.

[16] Izabel Cvetkovic, Valeria Rosenberg, and Eva Bittner. Conversational agent as a black hat: Can criticising improve idea generation? 2023.

[17] Kenneth E Iverson. Notation as a tool of thought. In *ACM Turing award lectures*, page 1979. 2007.

[18] Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walquer H Hill, David R Krathwohl, et al. Taxonomy of educational objetives: the classification of educational goals: handbook I: cognitive domain. Technical report, New York, US: D. Mckay, 1956.

[19] Matthias Felleisen, Robert Bruce Findler, Matthew Flatt, and Shriram Krishnamurthi. *How to design programs: an introduction to programming and computing*. MIT Press, 2018.

[20] Mark Guzdial. *Learner-centered design of computing education: Research on computing for everyone*. Morgan & Claypool Publishers, 2015.

[21] Gavriel Salomon. AI in reverse: Computer tools that turn cognitive. *Journal of educational computing research*, 4(2):123–139, 1988.

[22] Charles W Kneupper. Teaching argument: An introduction to the Toulmin model. *College Composition and Communication*, 29(3):237–241, 1978.

[23] Toshio Mochizuki, Toshihisa Nishimori, Mio Tsubakimoto, Hiroki Oura, Tomomi Sato, Henrik Johansson, Jun Nakahara, and Yuhei Yamauchi. Development of software to support argumentative reading and writing by means of creating a graphic organizer from an electronic text. *Educational Technology Research and Development*, 67:1197–1230, 2019.

[24] Martin Davies. Concept mapping, mind mapping and argument mapping: what are the differences and do they matter? *Higher education*, 62:279–301, 2011.

[25] Na Sun, Chien Wen (Tina) Yuan, Mary Beth Rosson, Yu Wu, and Jack M. Carroll. Critical thinking in collaboration: Talk less, perceive more. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, 2017.

[26] Sunok Lee, Dasom Choi, Minha Lee, Jonghak Choi, and Sangsu Lee. Fostering Youth's Critical Thinking Competency About AI through Exhibition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 2023.

[27] Adrian Holzer, Sten Govaerts, Samuel Bendahan, and Denis Gillet. Towards mobile blended interaction fostering critical thinking. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '15, page 735–742, New York, NY, USA, 2015. Association for Computing Machinery.

[28] Nathaniel Barr, Gordon Pennycook, Jennifer A Stolz, and Jonathan A Fugelsang. The brain in your pocket: Evidence that smartphones are used to supplant thinking. *Computers in Human Behavior*, 48:473–480, 2015.

[29] Raafat George Saadé, Danielle Morin, and Jennifer DE Thomas. Critical thinking in e-learning environments. *Computers in Human Behavior*, 28(5):1608–1617, 2012.

[30] Jiahao Yu, Xingwei Lin, and Xinyu Xing. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023.

[31] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*, 2023.

[32] Susmit Jha, Sumit Kumar Jha, Patrick Lincoln, Nathaniel D Bastian, Alvaro Velasquez, and Sandeep Neema. Dehallucinating large language models using formal methods guided iterative prompting. In *2023 IEEE International Conference on Assured Autonomy (ICAA)*, pages 149–152. IEEE, 2023.

[33] Andrew D. Gordon, Carina Negreanu, José Cambronero, Rasika Chakravarthy, Ian Drosos, Hao Fang, Bhaskar Mitra, Hannah Richardson, Advait Sarkar, Stephanie Simmons, Jack Williams, and Ben Zorn. Co-audit: tools to help humans double-check AI-generated content. *Proceedings of the 14th annual workshop on the intersection of HCI and PL (PLATEAU 2024)*, 5 2024.

[34] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.

[35] Advait Sarkar. Should Computers Be Easy To Use? Questioning the Doctrine of Simplicity in User Interface Design. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA, 2023. Association for Computing Machinery.

[36] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *arXiv preprint arXiv:2309.01029*, 2023.

[37] Advait Sarkar, Xiaotong (Tone) Xu, Neil Toronto, Ian Drosos, and Christian Poelitz. When Copilot Becomes Autopilot: Generative AI's Critical Risk to Knowledge Work and a Critical Solution. In *Proceedings of the Annual Conference of the European Spreadsheet Risks Interest Group (EuSpRIG 2024)*, 2024.

[38] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*, pages 3–10. IEEE, 2013.

[39] Elaine Simpson RN and Mary Courtney RN. Critical thinking in nursing education: Literature review. *International journal of nursing practice*, 8(2):89–98, 2002.

[40] Robert J Sternberg and Diane F Halpern. *Critical thinking in psychology*. Cambridge University Press, 2020.