AI SHOULD CHALLENGE, NOT OBEY *

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

ABSTRACT

Let’s transform our robot secretaries into Socratic gadflies.

1

“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 canonized great men of America are seldom unambiguously so. It aims to instill in students that rare and nebulous skill of critical thinking.

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

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 [1] 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” [2]. 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 output for accuracy and usefulness. It involves editorial decisions that demand creativity, expertise, intent, and critical thinking.

However, 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. This vision pervades AI interaction metaphors, such as Cypher’s “watch what I do” and Lieberman’s “your wish is my command”. 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.

2 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.

*Preprint of an article to appear in Communications of the ACM (CACM). The published version of the article may differ.
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 [3] to stimulate our thinking.

The idea of AI as provocateur complements, yet challenges, current frameworks of “human-AI collaboration”, which situate AI within knowledge workflows (notwithstanding objections to the term [4]). Human-AI collaborations can be categorised by how often the human (versus the AI) initiates an action [5], or whether human or AI takes on a supervisory role [6]. AI can play roles such as “coordinator”, “creator”, “perfectionist”, and “doer” [7] and “friend”, “collaborator”, “student”, “manager” [8]. Researchers have called for metacognitive support in AI tools [9], and to “educate people to be critical information seekers and users” [10]. 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” [11] framework distinguishes roles for critical thinking conversations, such as information gathering (white hat), evaluation and caution (black hat), etc. “Black hat” conversational agents, for example, lead to higher quality ideas in design thinking [12]. 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 [13]. 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?

3 Teaching critical thinking

The definition of “critical thinking” is debated. An influential perspective comes from Bloom and colleagues [14], 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 [15], and in Learner-Centered Design for Computing Education [16].

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 [17].

The Toulmin model decomposes arguments into parts like data, warrants, backing, qualifiers, claims, and their relationships [18]. Software implementations of this model help students construct more argumentative essays [19]. Similarly, “argument mapping” arranges claims, objections, and evidence in a hierarchy that aids in evaluating the strengths and weaknesses of an argument [20], and software implementations help learners [21].

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.

4 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 [2]. 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 [22]. Tools for preventing misinformation, such as Carl Sagan’s ‘Baloney Detection Kit’, can significantly impact user beliefs [23].
individuals are less inclined to engage in strenuous reasoning, they let technology take over cognitive tasks passively [24]. Conversely, the more interactive the technology, the more it is perceived to contribute to critical thinking [25].

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 [26], self-repair [27], and formal verification methods [28] can be integrated into the development and interaction loop to improve correctness. Language models can be designed to cite verifiable source text [29]. Beyond “correctness”, these techniques could also support critical thinking. A system error, if surfaced appropriately as a “cognitive glitch” [30], 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 [31]. Design questions include what kind of provocations, how many, and how often to show in particular contexts. These mirror longstanding questions in AI explanation [32], but as provocations are different, so the answers are likely to be.

Critical thinking is well-defined within certain disciplines, such as history [1], nursing [33], and psychology [34], 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.

5 Conclusions

“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.

References


AI Should Challenge, Not Obey (Sarkar, to appear in CACM 2024)


