Artificial Intelligence as a Tool for Thought

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

This is the author's annotated transcript of the following talk, containing additional commentary and references, indicated in red. Minor differences may remain between the transcript and the recorded talk.

> Advait Sarkar. How to stop AI from killing your critical thinking. TEDAI Vienna, September 2025.

ABSTRACT

Many AI tools focus on solving specific tasks like content generation or process automation. Though useful and powerful, these systems may affect how we think, learn, build skills, and develop expertise. We imagine how AI might help people to think better, so that: As well as getting the job done, it helps us better understand the job. As well as creating content, it helps us think more critically and with more insight. As well as seeking efficiency, it helps us create outcomes that are of higher quality because they are the product of better answers from better questions. As well as augmenting individual tasks, it augments collective workflows. As well as automating known processes, it helps organisations explore the unknown. We share principles for supporting cognition in any user experience, as well as new technologies that show what it means to support better thinking through AI.

Version: 13/11/2025

1 Talk transcript

1.1 Introduction: the effects of AI on thought

I'm here today to talk about thinking for yourself. And I must admit I did use AI to help me think about it. The irony is not lost on me.

But the way I did so is not by using AI as an assistant to help me prepare this talk faster. Rather, I used AI as a tool for thought. And by the end of this talk, I will have explained what I mean by that, why it is important, and given you a glimpse of how it might work.

But first, I need to set the scene.

Let's look at a day in the life of the 21st century knowledge worker.

I arrive at my office and look at my inbox full of emails. Ah. Let's summarise it.

I'm struggling to figure out how to respond here, so let's get AI to write a response.

Next, I need to write a report. But I'm struck by the blank page problem. I know, I'll drop in some resources and get an AI draft. Looks good to me! By the way, writer's block used to be staring at a blank page. Now it's staring at a page that AI filled out for me and wondering if I agree with it. I've become a professional validator of a robot's opinions.

I've got some data to analyse. Maybe AI can analyse this data for me? Probably correct.

OK, I've got to make a deck as well. You know the drill.

Oh, I was supposed to prototype something as well. OK, let me vibe code something.

Alright, all this looks good. Let's go.

This isn't a vision of the future. This is a completely plausible, if slightly exaggerated picture, of the world of knowledge work today.[1] Welcome to the age of outsourced reason, where the knowledge worker no longer engages with the materials of their craft. We've become intellectual tourists in our own work. We visit ideas. We don't inhabit them.

Our relationship to our work is entirely intermediated by AI. Some might say, alienated. We've heard that story before. What's wrong with this picture? For one thing, it is only one step removed from this, which is important, but that's a different talk.

What I want to focus on today is that using AI in this way can have profound implications for human thought.

Consider creativity. On an individual level, we might think that AI is a creativity boost, giving us rapid access to new ideas. But numerous studies have shown, that on a collective level, knowledge workers using AI assistance produce a smaller range of ideas than a group working manually.[2] We've created a hive mind, except the hive is really boring and keeps suggesting the same five ideas.

Consider critical thinking. We surveyed knowledge workers about their use of AI. They reported that they put less effort into critical thinking when working with AI than when working manually. And this effect was greater when workers have greater confidence in AI, and less confidence in themselves.[3]

Consider memory. When people rely on AI to write for them, they remember less of what they wrote. And when they read AI-generated summaries, it is hardly surprising that they remember less than if they had read the document.[4]

And finally, consider metacognition, which is the ability to think about your own thinking process. Working with AI requires significant metacognitive reasoning about your task goals, decomposing the task, the applicability of GenAI, your ability to evaluate the output.[5] These are things which are built in to the process of working directly with a material, and which become problematic when that material engagement becomes intermediated. Basically, we've become middle managers for our own thoughts.

So what's the score? We have fewer ideas, we think about them less critically, we remember them less well, and we have a harder time doing it. Taken together, we can see that AI-assisted workflows can have profound effects on human thinking.

And this extends even to seemingly trivial, mundane tasks, because these everyday opportunities for exercising our creativity, our critical thinking, and our memory, are essential for protecting our cognitive musculature, and allow us to rise to the occasion when an exceptionally complex task comes our way.[6] Studies show that when we don't use our brains, they get worse at brain things. Nobel Prize committee, please hold your applause.

Is this the cost of progress? We've solved the problem of having to think. Unfortunately, thinking wasn't actually a problem. It's like we invented a cure for exercise and then wondered why we're out of breath all the time.

[1] For examples of studies supporting the rapid and widespread adoption of AI at the individual and firm level, see, e.g. [Bick et al., 2024] and [OECD/BCG/INSEAD, 2025].

[2] This phenomenon, termed "mechanised convergence", was first introduced in [Sarkar, 2023] (Section 5.3), and expanded upon with more supporting evidence in [Sarkar, 2024b] (Section 2). A representative example of a study that supports the theory of mechanised convergence is [Dell'Acqua et al., 2023].

[3] Our survey: [Lee et al., 2025]. See also a related survey by [Gerlich, 2025].

[4] Neuroscientific explanation connecting cognitive offloading with decrease in IQ, focusing on memory [Oakley et al., 2025]. Study of LLM use in secondary schools showing lower retention and comprehension in LLM-only condition [Kreijkes et al., 2025]. See also memory results from [Kosmyna et al., 2025], but care must be taken when interpreting these results.

[5] A review that synthesises much available evidence into a single picture of the metacognitive demands created by GenAI: [Tankelevitch et al., 2024]

[6] The classic text on this phenomenon as it pertains to human-machine interaction is [Bainbridge, 1983], which assimilates multiple studies from the era to advance its argument. There have been many studies of cognitive offloading to contemporary digital technologies, e.g. smartphones [Barr et al., 2015]. See [Oakley et al., 2025] for the consequences of long-term underuse of the brain's declarative and procedural memory systems, associated with the use of GenAI.

It doesn't have to be this way. Beyond AI as an assistant, I believe that AI should be a tool for thought. AI should challenge, not obey.[7] And I believe that right at this moment, we are at a critical juncture, where the world of work is poised to be transformed by generative AI, and we must act now to shape and drive that transformation towards humanistic values. Of these two diverging roads, we must take the one less travelled.

Beyond getting the job done, a tool for thought helps us better understand the job. Beyond getting it done faster, it helps us get it done better. Beyond getting us to the right answers, a tool for thought helps us ask the right questions. Beyond automating known processes, it helps us explore the unknown.

What does this look like? What I'm about to show you is a prototype developed by my colleagues and me at the Tools for Thought team at Microsoft Research in Cambridge.[8] Now please bear in mind that this is a live research prototype. It's not a product, and it is just one of a series of explorations that our team is conducting to study how different modes of working with AI can enhance human thought.

1.2 Demo: a prototype tool for thought

So let's look at a fictitious example.

Clara and her colleagues run a company that sells bottled beverages. They have just had a meeting to discuss a new industry report that seems to have some pretty important findings about consumer preferences for sustainable packaging. Clara's colleagues have asked her to write a proposal arguing for how the company ought to respond. So she needs to really get to grips with this report, understand its findings and its data, and how it fits into her business context.

She starts by loading some documents into her workspace: there's the meeting transcript to remind her what was discussed, there's a recent internal report from her own business, and of course, there's the industry report, which she opens.

She sees an overview of the document, along with section by section summaries. Except these aren't really just summaries. We think of them more as lenses: they're customisable micro-representations of the text that can emphasise what is most relevant to the task at hand. So in this case Clara selects the consumers lens.

She can select a section for deeper reading, in this case the first one. As she reads, she makes notes about her thoughts and highlights excerpts from the document.

As she reads, she also sees AI-generated commentary and critiques. We call these provocations. Here's a provocation that raises a potential opportunity, which she highlights and annotates.

Note how this process is a hybrid of completely manual reading and completely relying on AI to read for you. Clara still reads, but intentionally and strategically.

Now as Clara is working, she is building up an outline of her argument, manually, in this pane on the right. This outline is lightly structured, and allows her to sketch out the flow of her argument at a high level, while still retaining deep connections to, and being grounded in, the source documents.

As a result of which we can already generate a draft of the proposal. And Clara can do things here like add a heading to the outline to generate a paragraph.

But what I want to draw your attention to here, is that while this text is AI-generated, Clara has a completely different relationship to this text than if she had just dropped in some documents and said "write me a report". Because this text is deeply rooted in a cognitively effortful, but interactionally effortless thought process. It reflects Clara's decisions, Clara's judgements, Clara's unique personal, professional expertise.

She sees another provocation, this time in the outline. In this case, she decides that while the provocation is useful, she does not need to address it. Unlike typical AI suggestions, provocations are not meant to be applicable all the time. They are instead meant to stimulate your thinking about your work. Because if you

[7] This maxim comes from [Sarkar, 2024a], which introduces the notion of AI as provocateur and as a tool of thought.
[8] https://www.microsoft.com/en-us/research/project/tools-for-thought/

understand your work well enough, deeply enough to make the confident decision not to accept a piece of feedback, then the feedback process is still working as intended.

But we're not done yet. Clara has entirely new ways of interacting with this text because of generative AI. A really simple example is that she can just resize a paragraph to change its length.

She can can also rapidly test different versions of this text. For instance, in this paragraph she is wondering whether it would be more effective if it took a more inspirational or a more practical tone. So she selects one of these customisable dimensions, and previews a few alternatives, and selects one.

And at select, strategic points, indeed, she writes. As she writes, she sees provocations that rather than autocompleting her ideas, they raise alternatives, they identify fallacies, they offer counterarguments to help her strengthen and develop her own argument.

There's something you won't find anywhere in this interface, and that's a chat box. Clara's not having to chat with anything to do her work. Yet she is silently and appropriately assisted by her computer, as a computer, and not as an ersatz human.

To put it simply, we have gone from this, to this.

Throughout this process, Clara has been assisted, and yes, probably worked faster, because of AI. But she has also maintained direct material engagement at strategic points. She read the relevant portions of the document herself. She constructed her decisions and her argument herself. And ultimately, it can be said, she has written this document herself. Moreover, she worked better because of AI. AI provocations at every stage of the process kept her metacognitively engaged, always looking for critiques, alternatives, and lateral moves.

1.3 Conclusion: the importance of human thought

We have been studying the effects of tools like this, and the results are promising. You can demonstrably reintroduce critical thinking into AI-assisted workflows.[9] You can reverse the loss of creativity and enhance it instead.[10]. You can build powerful tools for memory that enable knowledge workers to read and write at speed, with greater intentionality, and remember it too.[11] It turns out, with the right principles of design, you can build tools that are the best of both worlds: applying the awesome speed and flexibility of this technology to protect and enhance human thought.

These are simple, general principles like ensuring that the tool preserves material engagement, offers productive resistance, and scaffolds metacognition. And while we've been primarily studying professional knowledge workers, we believe that these principles can extend to all aspects of AI use, including when we use it in our daily lives, our hobbies, and even in education.

I repeat: efficiency is not the aim of tools for thought. Better thinking is. But sometimes, you can have both. I used to think that there was no such thing as a free lunch in human thinking. This is so much better than a free lunch. This is a lunch that pays you to eat it.

I want to close with some thoughts on the values that we have in developing AI software. What if AI gets to the point where it can do a better job of thinking than humans? Why should we care so much about protecting and augmenting human thought? There's two reasons: first, there may always be ways of thinking that remain unique human strengths of which we may not even be aware. Second, perhaps more importantly, we take the position that the ability to think well is essential for human agency, and empowerment, and flourishing.

[9] See our study of the effects of provocations on AI-assisted spreadsheet workflows [Drosos et al., 2025]. Numerous studies have demonstrated the potential for "cognitive forcing functions" to reduce overreliance on AI [Buçinca et al., 2021]. A compelling early study found that simply asking questions instead of providing explanations improves logical discernment [Danry et al., 2023]. This line of work has a long history in education research by which we are inspired, an early demonstration can be found in [Salomon, 1988].

[10] See, e.g. [Drosos et al., 2025], where median outcome diversity increased with AI provocations (although the difference is not statistically significant). Other recent experiments suggest that outcome diversity can be increased with personalised and more contextual prompting strategies [Wan and Kalman, 2025, Ghods et al., 2025].

[11] Reicherts et al. (forthcoming, study pre-registration information: `https://osf.io/ntxs4/overview`).

This echoes an ancient question. People once asked: if writing, if books, if the internet can remember for us, does it matter that we cannot? People once asked: if maps can navigate for us, does it matter that we cannot? Now we ask: if machines can think for us, does it matter that we cannot? If machines can speak for us, grieve for us, pray for us, love for us, does it matter that we cannot? To me the answer is pretty obvious. When I began studying human-AI interaction 13 years ago, it was inconceivable to me that we would be asking these questions in my lifetime. But we are. And we must.

I'll leave you with this thought. What would you rather have: a tool that thinks for you, or a tool that makes you think?

References

[Bainbridge, 1983] Bainbridge, L. (1983). Ironies of automation. In *Analysis, design and evaluation of man–machine systems*, pages 129–135. Elsevier.

[Barr et al., 2015] Barr, N., Pennycook, G., Stolz, J. A., and Fugelsang, J. A. (2015). The brain in your pocket: Evidence that smartphones are used to supplant thinking. *Computers in Human Behavior*, 48:473–480.

[Bick et al., 2024] Bick, A., Blandin, A., and Deming, D. J. (2024). The rapid adoption of generative AI. Technical report, National Bureau of Economic Research.

[Buçinca et al., 2021] Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21.

[Danry et al., 2023] Danry, V., Pataranutaporn, P., Mao, Y., and Maes, P. (2023). 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.

[Dell'Acqua et al., 2023] Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

[Drosos et al., 2025] Drosos, I., Sarkar, A., Toronto, N., et al. (2025). "it makes you think": Provocations help restore critical thinking to ai-assisted knowledge work. *arXiv preprint arXiv:2501.17247*.

[Gerlich, 2025] Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, 15(1).

[Ghods et al., 2025] Ghods, K., Liu, P., Labrou, K., MacDonald, K., Menon, A., and Wu, A. (2025). Evidence against llm homogenization in creative writing.

[Kosmyna et al., 2025] Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., and Maes, P. (2025). Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*.

[Kreijkes et al., 2025] Kreijkes, P., Kewenig, V., Kuvalja, M., Lee, M., Vitello, S., Hofman, J., Sellen, A., Rintel, S., Goldstein, D. G., Rothschild, D. M., Tankelevitch, L., and Oates, T. (2025). Effects of llm use and note-taking on reading comprehension and memory: A randomised experiment in secondary schools. Working paper, SSRN. Available at SSRN, posted January 13, 2025.

[Lee et al., 2025] Lee, H.-P. H., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., and Wilson, N. (2025). The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

[Oakley et al., 2025] Oakley, B., Johnston, M., Chen, K.-Z., Jung, E., and Sejnowski, T. (2025). The memory paradox: Why our brains need knowledge in an age of ai. In *The Future of Artificial Intelligence: Economics, Society, Risks and Global Policy*. Springer Nature. Forthcoming.

[OECD/BCG/INSEAD, 2025] OECD/BCG/INSEAD (2025). The adoption of artificial intelligence in firms: New evidence for policymaking. Report, OECD Publishing, Paris.

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

[Sarkar, 2023] Sarkar, A. (2023). 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. ACM.

[Sarkar, 2024a] Sarkar, A. (2024a). AI Should Challenge, Not Obey. *Communications of the ACM*. Online First.

[Sarkar, 2024b] Sarkar, A. (2024b). Intention Is All You Need. In *Proceedings of the 35th Annual Conference of the Psychology of Programming Interest Group (PPIG 2024)*.

[Tankelevitch et al., 2024] Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., and Rintel, S. (2024). 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. Association for Computing Machinery.

[Wan and Kalman, 2025] Wan, Y. and Kalman, Y. M. (2025). Using generative ai personas increases collective diversity in human ideation.