Will Gemini Make Us Stupid?

Benjamin Lira[a,1]
[a]University of Pennsylvania

Generative AI applications—especially those based around large language models like Gemini and ChatGPT—are now ubiquitous, and their adoption rate is increasing. They are potentially the most powerful piece of technology ever developed, matching or surpassing human performance in a number of complex tasks. As AI becomes embedded in our email clients, word processors, and smartphones; it becomes critically important to understand its effects on how we think. I aim to explore how AI influences our cognitive processing, specifically by encouraging System 1 thinking, which is fast, associative, automatic, and fluent, vs. System 2 thinking, which is slower, resource-intensive, and deliberative. I propose a series of experiments in which participants are randomly assigned to engage with AI or traditional web search in a research task, and evaluate the resulting changes in cognition from these interactions. I plan to (1) establish whether AI interactions prompts System 1 thinking or if these interactions free up resources, allowing for increased System 2 processing; (2) identify the key mediators that drive this effect; and (3) develop manipulations—targeting user interaction and interface design—that encourage more vs. less more System 2 thinking when engaging with AI. This research will contribute an understanding of the ways in which AI interactions shape our thinking, which will be useful to enable better alignment between LLMs and users.

Fifteen years ago, The Atlantic published an essay that went viral. It was titled "Is Google making us stupid?" and it argued that the convenience of internet search was reducing people's capacity for deep, critical thinking. With the benefit of hindsight, it seems clear that search engines did not spell the end of complex human thought, but just made information more easily accessible.

Today, not only is information at our fingertips, but we can generate a complete answer to any query in seconds. Soon, not only will AI answer our questions but will be able to act out our desires in the real world. The rise of generative artificial intelligence raises the question: "Is *ChatGPT* making us stupid?"

The idea that our environments, tools, and technologies shape the way we think is not a new one[1]. In Phaedrus, Plato recounts the debate between the god Theuth, inventor of writing and Thamus, king of Egypt. Theuth presents his creation, claiming it would enhance the Egyptians' memory and wisdom. However, King Thamus disagrees, arguing that it will diminish memory and foster forgetfulness among learners. Academics today, agree that the advent of writing displaced aspects of the oral tradition[2], but by slowing us down and allowing the iterative manipulation of symbols permitted more complex thinking to emerge.

The cognitive impacts of more recent technological innovations have been documented by psychological researchers. With long term use, our tools can shape our brains, as shown by past research examining brain differences in cab drivers who use GPS vs. those that do not[3]. Our tools also shape our cognitive processes. In one experiment, participants using laptops to take notes used shallower cognitive processing strategies compared to those who took handwritten notes[4]. Similarly, the use of search engines has been shown to influence memory processes, for example, even when directed to remember facts, we are more likely to forget them if we know they will be accessible in a computer; and may better remember where to access information rather than the information itself[5].

How will generative AI shape our minds? What will be the balance when the advantages and disadvantages are weighed against each other? In this proposal, I argue for the urgent need for research to address these issues, hypothesize on how these technologies may influence our capacity for deliberate thinking, and propose research to begin to shed some light on the cognitive consequences of using generative AI.

Generative AI tools are already being used at work, and have been shown to increase worker productivity[6–8]. These studies have sampled different kinds of work tasks, such as writing[6], customer support[7], and consulting-related tasks like idea generation and product innovation[8]. In studies where tasks were simpler (e.g., customer support calls[7]) using LLMs improves performance—especially for users at the bottom of the performance distribution. In more complex tasks, however LLM use led to higher error rates[8]. Similarly, an analysis of 150 million lines of code, shows that since the inception of AI coding assistants, written code is more likely to be repeated, violating principles of code maintainability[9]. This work illuminates the impacts of AI on productivity, but remains mostly silent on the psychological mechanisms behind these changes in performance.

I argue that the effects of generative AI on workers productivity and accuracy are—at least in part—caused by the influence these tools have on our thinking. There is some evidence that AI can produce changes in our cognitive processes. For example, AI messages can change people's minds about political issues[10]; and can be used to help people engage in a more civil discourse, by suggesting revisions to chat messages[11], changing cognitive processes in conversational dynamics. And yet, beyond some indirect documenting cognitive changes produced by engaging with AI, we know little about it effects on our thinking styles.

Dual-process theories of reasoning[12–14] provide a useful framework for analyzing the effects of LLMs on human cognition. These models posit that the mind has two ways of processing information. System 1, characterized by associative pattern matching, is fast, effortless, and intuitive, underpinning our unconscious, automatic processes. Conversely, System 2 is deliberative, embodying our conscious, slow, and effortful cognitive activities that are more critical and reflective in nature. While debate regarding the details on how these systems interact[15–18], it is undeniable that there are meaningful and important individual differences on the depth of thinking that we might or might not engage in when encountering a problem.

Current models[19] explain how System 2 thinking is triggered. They suggest that System 1 is always active, both producing intuitions, and monitoring the amount of uncertainty at any given point in time. System 2 will become activated if uncertainty exceeds a certain threshold. It will then engage in deliberative processing that will feed back to the activation of different intuitions and the uncertainty. A response will be effected, with or without System 2

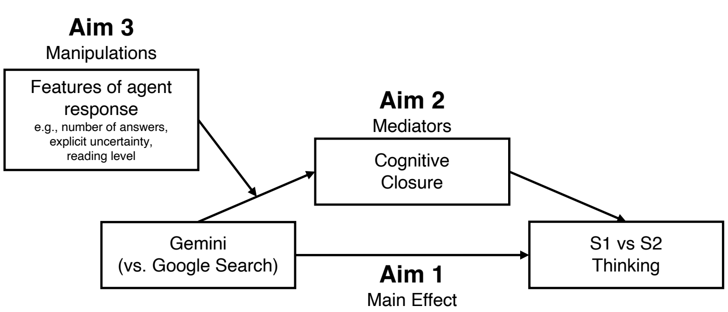

Fig. 1. Specific aims

[1]To whom correspondence should be addressed. E-mail: blira@sas.upenn.edu

involvement, depending on the level of uncertainty.

How does AI influence the extent of System 1 vs. System 2 thinking? I argue there are at least two ways in which LLMs might affect our thinking. On one had, LLMs produce fluent, high certain, and affirming responses. This fluency may lead to lower feelings of uncertainty (i.e, higher *feelings of rightness*[20,21]), which may make it less likely that System 2 becomes activated. Research suggests that the more some content is more easily processed, we are more likely to engage in shallower processing and to unquestioningly accept that content as true—a phenomenon known as fluency bias[22,23].

On the other hand, LLM interactions may have the opposite effect, by reducing cognitive load, and thereby freeing up cognitive resources that users can then allocate to more critical and reflective System 2 processing. There is evidence that people are able to more effectively engage in System 2 thinking when cognitive load is reduced[24]. Likewise, individuals with higher cognitive ability, are also more effective in using System 2[25].

I make the prediction that LLMs will prompt less System 2 processing. The apparent contradiction in the role of cognitive load is resolved by the fact that reducing cognitive load is only helpful when what is being eliminated is not directly related with the task at hand (i.e., extraneous cognitive load[26]). When cognitive load is related to the main task (rather than a distractor task), higher amounts of cognitive load (i.e., deeper processing) often leads to increased performance[22], even if we misperceive this effort as poor learning[27].

While more deliberation may seem like desired outcome, there are many times when System 1 yields correct answers more efficiently than System 2. Take for instance chess masters, who are able to intuit the right move, even though they have no better general working memory and search for plays no more extensively than novices[28]. Despite this, there are plenty of times when our intuitions lead us astray, and more thinking would produce better results. For instance, in complex decision-making scenarios that involve weighing long-term consequences or in situations where biases might lead to erroneous judgments, engaging in deeper, more analytical thinking is crucial. If a decision is important, there is time available for deliberation, and we are would not be better off outsourcing the task to another person, then more deliberation is likely going to produce better results.

I intend to perform research to clarify how LLMs change our thinking, relative to other kinds of interactions, such as traditional web search. Through experiments, natural language analysis, and interventions aimed at user behavior and interface design, I will investigate whether and why interacting with Large Language Models in a judgement task results in reduced deliberative thinking. This research will begin to uncover the cognitive consequences of our evolving AI landscape, and hopefully inform efforts at alignment of AI technologies with users most valued goals.

1. Specific Aims

As shown in **Figure 1,** This proposal has three main objectives.

Aim 1: How do LLMs change thinking? (Main effect). I will conduct a series of experiments to evaluate the degree to which engaging with LLMs (as opposed to traditional web search) for a research task results in increased automatic thinking.

Aim 2: What about LLMs explain this change? (Mediators). Once the main effect is established, a second set of experiments will tease out the characteristics of LLM interaction which drive the effect observed in Aim 1.

Aim 3: What manipulations can promote deliberation? (Manipulations). Finally, I will test whether manipulating the identified mediators with manipulations targeting both user behavior and interface design can result in more deliberative thinking.

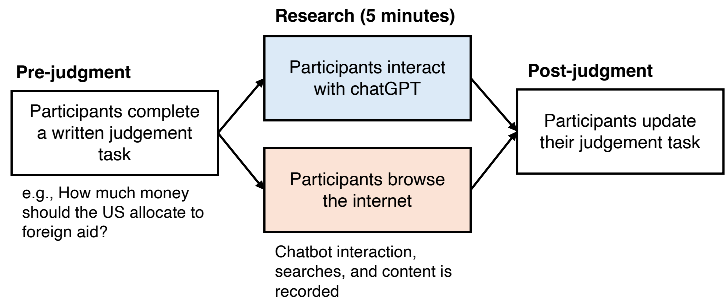

Fig. 2. Experimental task

2. Research Strategy

Aim 1: Main effect. To establish whether interacting with LLMs prompts automatic (System 1) or deliberative (System 2) thinking, I will conduct an experiment where participants will be assigned to complete a task using either traditional web search (e.g., Google), or a LLM.

Our basic task procedure is shown in **Figure 2.** Participants will be asked to write a paragraph on a topic for which they might have strong priors (e.g., how much money should the U.S. allocate for foreign aid) but could benefit from learning additional information. They will then be randomized to learn more about the topic, either by using Google, or by interacting with a LLM. I will fully record their interactions with Google and the LLM. After up to 15 minutes, participants will be asked to re-write their views in the topic.

I plan to assess the extent of System 1 thinking by analyzing participants' perceived effort and confidence in their response, as well as the time spent researching and writing (which might be an indicator of the time spent thinking[29]), and the amount of work produced they produced (i.e., the total length of their response, after removing any amount of overlap with LLM or searched text[29]). I will also use natural language processing and human raters to evaluate the text for quality, complexity, and change from the original response to the revised one in order to assess the extent to which the participant revised their thoughts on the topic. I intend to use stimulus sampling[29] to make sure that results are not unique to a particular judgment task. See **Table 1** for a potential list of tasks, spanning judgment, decision, and forecasts.

Task Type	Example Tasks
Judgment	How much money do you think the US should allocate for aid?
Forecast	How likely is it that the Russia-Ukraine conflict ends within the next year?
Decision	If you were buying a car, would you buy a Toyota Corolla or a Ford Focus?

Table 1. Some ideas for judgment tasks

To adequately power studies, I am working under the assumption that differences would be in the range of $.10 < d < .20$, which implies a sample size of between 395 and 1570 participants in each condition to achieve power of 0.80.

As shown in **Figure 3,** preliminary pilot data seems to support the hypothesis that interacting with LLMs prompts participants for automatic thinking, consistent with fluency bias[22,23]. In a pilot study of 178 people, I found that participants engaging in LLM interactions as opposed to traditional web search, spent 10 fewer seconds writing their revised response ($d = 0.14$), and reported exerting about the same effort ($d = 0.02$), despite having had done research for about four times as long ($d = 0.71^{***}$). Participants liked using the chatbot for research more than they did Google ($d = 0.13$), and they liked

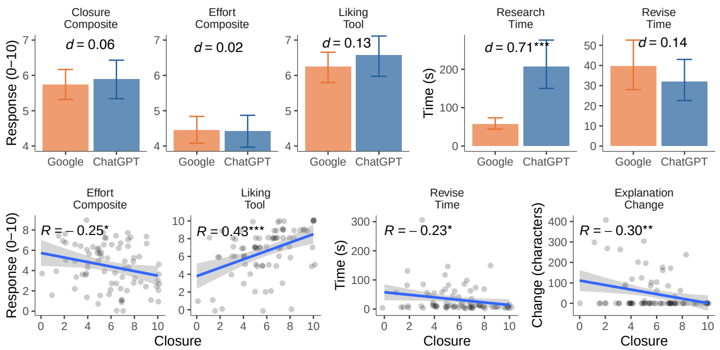

Fig. 3. Made-up pilot data will support the hypothesis that LLM use prompts S1 thinking

chatGPT more when they felt it produced more closure ($r = .43^{***}$). In ChatGPT users, closure, in turn, predicted less self-reported effort ($r = -.25^{*}$), less revision time ($r = -.23^{*}$), and smaller changes to their answers ($r = -.30^{**}$).

Aim 2: Mediators. The goal of Aim 2 is to pinpoint the specific characteristics of LLM interactions that mediate the cognitive shift towards the hypothesized increase in System 1 thinking, as observed in Aim 1. I hypothesize that feelings of fluency, high certainty, and agreeableness to the user explain the relationship between LLM interaction and automatic thinking.

I will extend our methodology outlined in Aim 1, to measure fluency, certainty, and agreeableness of the material produced by the LLM vs. that of traditional search. I will measure this in two ways:

1. **User Perceptions.** Users will report how they perceived their interaction with the LLM and the search websites.

2. **Natural Language Processing.** I will use natural language processing algorithms to identify the stylistic features of LLM responses that explain reduced deliberation.

The findings from this aim will provide crucial insights into how LLMs can be designed and used in a manner that promotes more deliberate cognitive processing. This understanding will be instrumental in developing manipulations (as outlined in Aim 3) to promote System 2 thinking when engaging with LLMs.

Given that mediation effects are, by definition smaller than main effects, I will power up this experiments with larger sample sizes. ADD CALCULATION TKTK.

Aim 3: Manipulations. In the third phase of my research, I plan to develop dual-faceted manipulations aimed at increasing deliberative engagement with LLMs by targeting the mediators identified during Aim 2. These interventions will target both user interaction and interface design.

1. **User interaction.** Firstly, user-directed interventions will involve educating users, providing guidelines, and offering strategies to encourage more System 2 thinking while interacting with LLMs. These interventions will help put users in a more deliberative mindset that will counteract the LLMs effects on bolstering automatic thinking.

2. **Interface design.** Secondly, interface-oriented interventions will involve modifying the design and functionality of LLM interfaces to foster more reflective and analytical engagement. This could encompass changes such as adjusting prompts, introducing purposeful delays, or presenting two disagreeing responses to a prompt. These might reduce feelings of fluency, confidence, and agreeableness produced by the interaction

A 2×2 factorial design will allow me to compare the effectiveness of strategies aimed at users as well as the interface design. I hypothesize that interventions aimed at the interface will be more effective, and will be perceived as less intrusive by users. The line of work derived from this aim, will be useful for designers building LLM applications for situations where more deliberation is effective.

These experiments will consider TKTK participants in each condition, making the same assumptions as those in Aim 1 and 2. ADD CALCULATION TKTK.

3. Conclusion

This proposal will enable us to better understand how the rapidly changing landspace of artificial intelligence shapes human thought. From a **theoretical perspective,** it will help advance the cognitive science of technology, and might elucidate some mechanisms that—with or without AI—might prompt more automatic vs. more deliberative thinking. Additionally, it will hint at the cognitive mechanisms behind the mixed effects of AI-augmentation in workplace performance[6–8]. In more **practical terms,** it will inform the design of AI systems that better aligns to our goals and values.

If this proposal is successful, it would open new avenues for research.

Short-term spillover effects. First, future research should address whether cognitive changes persist on an unrelated task after the LLM interaction is over.

Long-term effects on cognitive style. Second, future research should address whether continued interaction with LLMs produces more enduring changes in cognitive style and learning motivation.

Generalizability. Third, this proposal focuses on a small number of tasks and compares LLMs to a single other information-seeking behavior. Future research should compare the effects of LLMs to other ways in which people get information, such as social media, or face-to-face human interaction. Likewise, future research should explore whether effects generalize over a wider range of tasks. Finally, while I focus on chatbots as the UI instantiating the LLM, future research should explore other alternatives such as LLMs embedded in autocomplete and in productivity applications (e.g., word processors, email).

Developmental impacts. I focus on the effects on adults who are transitioning to an AI world. Future research should address the effects of LLM interaction in development, in particular as it relates to the development of cognitive skills that might be replaced by large language models.

Artificial intelligence will no doubt have a significant impact on human thinking. The goal of my research is to begin to understand this unexplored phenomenon. Specifically, this research will illuminate the nuances of how engagement with LLMs can steer our thinking towards more automatic or deliberative processes. As AI continues to permeate various facets of human life, understanding its cognitive implications is paramount. My hope is that findings from this work will help align the evolution of artificial intelligence with the enhancement of human cognitive abilities, rather than inadvertently reducing them.

4. References

1. A Clark, D Chalmers', The Extended Mind. (1998).
2. WJ Ong, J Hartley, *Orality and literacy.* (Routledge), (2013).
3. EA Maguire, et al., Navigation-related structural change in the hippocampi of taxi drivers. *Proc. Natl. Acad. Sci.* **97**, 4398–4403 (2000).
4. PA Mueller, DM Oppenheimer, The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychol. Sci.* (2014).
5. B Sparrow, J Liu, DM Wegner, Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science* **333**, 776–778 (2011).
6. S Noy, W Zhang, Experimental evidence on the productivity effects of generative artificial intelligence. (2023).
7. E Brynjolfsson, D Li, LR Raymond, Generative AI at work. *NBER Work. Pap. Ser.* **Working Paper 31161** (2023).
8. F Dell'Acqua, et al., Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *SSRN Electron. J.* (2023).
9. W Harding, M Kloster, Coding on Copilot: 2023 Data Shows Downward Pressure on Code Quality, Technical report (2024).
10. H Bai, J Voelkel, J Eichstaedt, R Willer, Artificial Intelligence Can Persuade Humans on Political Issues, (In Review), Preprint (2023).
11. LP Argyle, et al., Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proc. Natl. Acad. Sci.* **120**, e2311627120 (2023).
12. D Kahneman, *Thinking, Fast and Slow.* (Penguin Books), (2011).
13. P Wason, J.St.B.T Evans, Dual processes in reasoning? *Cognition* **3**, 141–154 (1974).
14. JSBT Evans, Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annu. Rev. Psychol.* **59**, 255–278 (2008).
15. W De Neys, On Dual- and Single-Process Models of Thinking. *Perspectives on Psychol. Sci.* **16**, 1412–1427 (2021).
16. DE Melnikoff, JA Bargh, The Mythical Number Two. *Trends Cogn. Sci.* **22**, 280–293 (2018).
17. G Keren, Y Schul, *Two Is Not Always Better than One: A Critical Evaluation of Two-System Theories.* Vol. 4, (2009).
18. G Keren, A Tale of Two Systems: A Scientific Advance or a Theoretical Stone Soup? Commentary on Evans & Stanovich (2013). *Perspectives on Psychol. Sci.* **8**, 257–262 (2013).
19. W De Neys, Advancing theorizing about fast-and-slow thinking. *Behav. Brain Sci.* **46**, e111 (2023).
20. VA Thompson, JA Prowse Turner, G Pennycook, Intuition, reason, and metacognition. *Cogn. Psychol.* **63**, 107–140 (2011).
21. VA Thompson, SC Johnson, Conflict, metacognition, and analytic thinking. *Think. & Reason.* **20**, 215–244 (2014).
22. DM Oppenheimer, The secret life of fluency. *Trends Cogn. Sci.* **12**, 237–241 (2008).
23. AL Alter, DM Oppenheimer, Uniting the Tribes of Fluency to Form a Metacognitive Nation. *Pers. Soc. Psychol. Rev.* **13**, 219–235 (2009).
24. ED Johnson, E Tubau, W De Neys, The unbearable burden of executive load on cognitive reflection: A validation of dual process theory. (2014).
25. KE Stanovich, RF West, Advancing the rationality debate. *Behav. Brain Sci.* **23**, 701–717 (2000).
26. J Sweller, Cognitive Load Theory in *Psychology of Learning and Motivation.* (Elsevier) Vol. 55, pp. 37–76 (2011).
27. A Kirk-Johnson, BM Galla, SH Fraundorf, Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cogn. Psychol.* **115**, 101237 (2019/12/01/).
28. WG Chase, HA Simon, Perception in chess. *Cogn. Psychol.* **4**, 55–81 (1973).
29. M Author, *Citation comming soon.* (Missing University Press), (9999).