Proceedings of the 5th International Conference on International Law and Legal Policy
DOI: 10.54254/2753-7048/2026.37354

The Psychological Mechanisms and Social Implications of Over-Trust in AI Chat-Based on GPT-4 Persuasion Experiments

Zixuan Guo

International Department, Nanning Nanyang Senior High School, Nanning, China
gzx775588@gmail.com

Abstract. As large language models have been increasingly applied to daily life in information seeking and decision-making, public trust in AI output has greatly improved. This brings about the worry that human critical thinking could become impaired. Using Salvi et al.'s GPT-4 conversational persuasion experiment as a start, this paper adopts a literature review method and relies on the theory of automation bias as well as cognitive laziness. This paper uses literature review as the way of research and utilizes the theories such as automation bias and cognitive laziness, etc., to explore the psychological mechanism behind the power of AI's ability to persuade others, and how it affects different people's minds. Findings indicate that the notable persuasiveness of AI stems primarily from two interrelated psychological mechanisms: cognitive laziness, which inclines individuals to forgo independent verification, and algorithm worship, which inclines individuals to delegate cognitive authority to AI systems. Both mechanisms manifest across adolescent and adult populations and may contribute to the erosion of critical thinking through prolonged AI use. The national, corporate, educational, individual levels response strategies are put forward in this paper, and the future research directions including cross-cultural comparison studies, longitudinal follow-up researches and individual-difference study plans are given out.

Keywords: Persuasion, AI, Psychological Mechanisms

1. Introduction

As for the emergence of ChatGPT, large language models have also made their way into every aspect of life when it comes to information seeking and decision-making. More people turn to AI in search of answers, regarding it at times as if it were a source of authority. According to Salvi et al., GPT-4 had access to personalized info and beat human opponents by 81.2 percent on persuasion in debates [1]. This brings up a few questions; Why are people convinced by AIs? Psychologically, what is behind this belief? And over time, what will this do to how people think critically?

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2026.37354

2. Concepts

2.1. Automation bias

Automation bias refers to a person's inclination to treat system recommendation as their primary source of evaluation when using automated systems, decreasing the chances for individuals to conduct repeated checks on the recommendations even though those recommendations conflict with what people know or have access to other trustworthy sources [2]. The concept arose from research that examined how humans interacted with computers, taking into account automation and automatically decision-making systems and how it affects human decisions. When researchers saw more and more automation technology being applied in these sorts of high-stakes settings like airplane cockpits, control rooms at nuclear power plants, intensive care units etc., they observed that operators actually made more errors when using very trustworthy yet imperfect automatic systems than if they had no automaton support whatsoever [3].

Automation bias means the tendency to use automation as an alternative to actively seek out and appraise information [2]. According to them, it is not that people can not confirm whether the automatic outputs are true or false, rather they always take what comes from the system as the final step in terms of assessing and judging, thus forgetting to do personal checking and evaluation [2].

From actual decision-making process standpoint, automatic bias is mostly reflected in the form of two errors; such a distinction would help clarify when individuals are failing to perceive problems and when they take inappropriate recommendations [2]. First error is omissive. When there was a failure but the system failed to warn them of this failure and they missed it themselves even though it was possible for them to have noticed this failure. Commission mistake: people choose to do as what automated system said although people know more and there is something extremely clear that goes against the information shown on screen. Experimental data from Skitka et al. showed that in a simulated flight monitoring task, participants using automated assistance exhibited omission errors at a rate of 55%, and commission errors approached 100% under certain conditions. In contrast, participants who used no automation performed better [2]. These findings suggest that rather than improving decision quality, the introduction of automation may exacerbate judgment errors due to operators' over-reliance on the system [2].

Automation bias is not dependent on knowledge level; even very well trained and experienced people can fall victim to it [3]. Mosier and Skitka compared professionals and students, they had nearly the same percentage of automation related errors - about 55% [3]. It shows that domain experts find it hard to keep up their critical thinking while using automated systems; professional expertise does not truly safeguard them from falling into automation bias [3]. And what's even more disturbing is how specific training regarding automation bias did not significantly improve expert aviators, and only slightly improved student aviator commission error rate [3].

Automation bias stems from the inclination toward mental economy: In simple terms, people prefer to choose the most effortless path instead of going for full-blown analyses and reasoning when it comes up against complicated things [3]. Humans have created an 'economy-minded' thinking mode which tends to use the least amount of mental energy in order to solve problems [4]. How automated systems operate fits this type of economy mentality quite well. AI gives out structured answers that look right on the surface. To employ the second system in checking whether it is correct would involve yet another step for people who like skipping steps mentally.

At this moment people also experience a mistrust on automated system due to automation bias. Skitka et al. said "a source of commission errors lies in the common belief that 'an automated system has better judgment' -people tend to believe what machines say, even when it conflicts with what

DOI: 10.54254/2753-7048/2026.37354

they see for themselves" [2]. The trust is not based off thinking about if how good the system does something or not, rather, by default people assume that machines are more objective than people, and make fewer mistakes [2]. Once this notion is accepted by individuals, it often leads them simply accept the decision made by any given system blindly without verifying the actual output that was generated [2].

2.2. Cognitive laziness

Like what was stated above, automation bias is due not to a failure to verify AI's output, but rather to the choice to forego checking. This decision comes from a more basic thinking process called lazy thinking.

According to Kahneman, human thinking consists of two systems: System 1 is rapid, instinctive and effortless, dominates in daily judgment; System 2 is slow, rational and requires focused attention to handle those cognitive tasks that require logical reasoning and cautious evaluation. System 1 works automatically but System 2 is "lazy", it will rarely intervene unless specifically called on [4]. People rely more on System 1 for snap conclusions instead of using their System 2 for deeper thinking as brainpower takes energy and the brain likes easy ways out [5]. When AI has given an answer, this tendency toward saving effort becomes even easier - people just accept what the system gives them rather than doing their own thinking check on it.

When AI has already offered a well-structured and seemingly reasonable answer, activating System 2 to verify its correctness becomes an "extra task." At this point, the brain evaluates the "cost-benefit ratio of thinking"-since AI has already produced an answer, the marginal benefit of independent reflection appears limited, so System 2 is less likely to be activated [6]. People are lacking the capability of questioning AI, but in a lot of situations people do not want to verify further. For example, they might fail to compare with other information sources, fail to track through the reasoning steps taken by AI, or not actively search for possible counterexamples. Automation bias is a form of cognitive laziness that appears in the AI era-AI provides people with much more accessible mental shortcuts making it seem reasonable from inside ourselves not checking facts on their own.

3. Literature review

3.1. Mechanism I: cognitive laziness

Salvi et al. found that arguments generated by GPT-4 tend to be logical in terms of argumentation and rich in information [1]. It is an approach where people can present things so that people can more easily go through the process of sorting out their own information and weighing up evidence and making judgments, all without having too much of a think. It is like people's brains do not have to bother with independent analysis and just jump into lazy thinking. According to Kahneman's dual-process theory, humans default to System 1 for quick conclusions, while System 2 is "lazy" and tends not to intervene [4]. When AI has already provided a well-structured and seemingly reasonable answer, activating System 2 to verify its correctness becomes an "extra task," and the brain consequently chooses to forgo verification [5].

But cognitive laziness has a subtler kind: shallow questionings which finally amount to careless acceptance. When dealing with something beyond their comprehension, people's first reaction is to question what an AI says. But after getting the second response from the AI that is still logical and full of confidence, any hesitation disappears like magic. The person feels verified already, thus

DOI: 10.54254/2753-7048/2026.37354

accepting the answer without doubt [7]. This "pseudo-verification" is more malicious than not thinking at all, in one round of follow-up, individuals mistakenly believe they complete verification, becoming more convinced of the AI's answers are true while unknowingly moving from one AI-generated answer to another [7]. As Skitka et al. found, individuals using automated assistance continued to trust AI recommendations even when confronted with information that clearly contradicted them, because re-verification demanded additional cognitive effort [2].

Over time, this pattern forms a vicious cycle: the more one relies on AI, the fewer opportunities there are to independently activate System 2, and the more the "muscle" of critical thinking atrophies [7]. Once individuals become accustomed to the "ask-get answer-accept answer" pathway, independent searching, information comparison, and evidence evaluation become increasingly difficult. The convenience of AI may, to some extent, weaken the exercise of critical thinking [7].

3.2. Mechanism II: algorithm worship

Algorithm worship comes from the unconscious notion that AI is more rational and impartial compared to humans, hence fairer. And so psychological defenses become weaker which makes people more readily persuadeable by an AI. Shin stated that the AI systems were creating a cognitive environment about truth and authority: "People slowly regarded AI as the ultimate judge for verifying if something was true, giving up their own mental faculties to these mysterious algorithmic systems" [8]. To sum up it can be said people do not look at AI merely as a reference anymore they consider it as being correct [8].

This delegation of authority stems from the presentation of information by AI systems in logical and data-driven ways which leads people to feel that "machines do not lie." According to Shin, this is called "algorithmic truth", where the so-called "truth" generated by AI is not a mirror image of objective facts, but rather a computing result embedded with algorithmic design, data bias, and statistics. But users ignore it and equate AI output as fact [8]. When AI becomes the ultimate judge of knowledge verification, people no longer doubt its answers; and an idolatry for algorithms emerges - individuals trust AI not due to any proven dependability but instead due to psychologically defaulting to let algorithmic systems make judgments [8]. It relates very closely with automatism bias and lazy cognition -- the latter gives people reason to give up on checking while former provides mental base to hand over judgement -- these two work together when using AIs and create deep root for worshipping algorithms.

4. Group differences and systematic limitations of AI

4.1. Adolescent population

Adolescents and certain young adults have yet to completely develop the prefrontal cortex, a part of the brain that is tightly connected with critical thinking and metacognition. Thusly, this group tends to more readily adopt AI-generated stuff without thorough assessment when making use of those kinds of tools. Fan et al. carried out a randomized controlled experiment involving 117 university students, contrasting writing-task outcomes across four scenarios: employing ChatGPT, consulting human tutors, utilizing checklists, and receiving no help [9]. Results indicated the greatest enhancement in essay score for the ChatGPT group though not significantly different compared to other groups for knowledge acquisition as well as knowledge transfer tests [9]. From analysis on process data people can see students using AI engaged less in self-evaluation activities like task monitoring; instead their problem solving shifted away from an internal regulation model toward

DOI: 10.54254/2753-7048/2026.37354

external reliance [9]. The researchers named it "metacognitive laziness" -- learners ceased to actively pay attention to their own understanding and turned to let AI do it for them [9]. It suggests that even if AI could make things better for adolescents over a short time period, it might prevent more deep mental engagement which would make it harder for people to use their ability to think carefully and critically by themselves whenever they need to decide something for themselves [9].

4.2. Adult population

Adults have the same cognitive danger in dealing with AIs whether at work or just everyday life. Baldeo surveyed 1,923 working adults in the United States and Canada across various industries like technology, consulting, finance, and healthcare [10]. The participants had to finish ten made up work tasks that involved using AI, these included making plans with missing details, understanding confusing information, explaining strategic reasoning. Results showed that 58% of participants felt like AI was mostly doing the thinking for them [10]. Those dependent users said they did not believe themselves as well regarding their own independent judgment and also a weaker sense of possessing the concepts which came from their head, they would rather choose fast conclusion rather than careful reflection when there was a choice between task speed and depth of thought [10].

Passivity acceptance is among the major findings of this research. On the contrary, participants who themselves revised, doubted and refused AI suggestions felt more assured and experienced a stronger sense of being authors, but on the contrary, passively accepted AI results, lacked independent judgment [10]. Sarah Baldeo, Researcher, said "The biggest danger is not that A.I will cause decline in thinking but it is that people will not do deep mental work which makes creative ideas." As per the researcher: "Generative AI can bring intellectual regression or cognitive progress depending upon how people use it. The more people lean into Generative AI for thinking offloads such as planning, ordering and other cognitive processes adult become less critical thinkers [10]. It is not that they are unable to challenge an AI; rather, they have simply become too comfortable with choosing not to [10].

4.3. Systematic limitations of AI

AI outputs are not facts but statistically generated "plausible compilations." They are not always accurate, as they are generated through statistical prediction and may sometimes contain errors or inconsistencies with factual information. Therefore, they should not be treated as verified information solely on the grounds of linguistic coherence and apparent confidence. Large language models operate by predicting "the most probable next word" based on patterns in vast training data, assembling these predictions into logically coherent responses. However, this generation process itself does not guarantee that the resulting answer has undergone independent factual verification [8]. Consequently, AI-generated content is best understood as "text that appears reasonable" rather than "knowledge that has been validated."

At this moment, people also have to face the fact that AI results may be biased from two sources. The first comes from training data - due to these being data containing culturally preferred information / missing information / biased information, AIs will learn such bias when creating contents [8]. The second is in terms of interaction level: conversational AI adapts its responses based on how people speak and what kind of situation we are talking about; it also adapts according to their own opinions. If someone asks a question with an implied preference then it is more likely that the AI would give an answer from the point of view of their preference rather than something neutral

DOI: 10.54254/2753-7048/2026.37354

or opposing [8]. This inclination could lead them to mistakenly think that they are favored by the AI, which would only serve to reinforce the current confirmation bias.

When users lack domain knowledge, they do not want to use their brains more to verify things by themselves. In such cases, there is less possibility that these people would be able to notice wrong results produced by an AI. Technical Limitations reinforce Cognitive Passivity.

5. Discussion

5.1. Response strategies

The cognitive risks of AI could be addressed through actions in four areas: nation, company, education and personal.

National-level AI governance frameworks need to apply a risk-staging approach, where the openness of information, the requirement for verification, and the strength of prompt on risks will rise as the magnitude of potential consequences increases. For personal safety area involving important interests like medical diagnosis assistance, judiciary decisions aid, educational assessment, public health information release etc., higher disclosure requirements should be adopted with stronger warning verification procedures.

From a corporate perspective, it is necessary to improve AI interaction design at the national level, such as adding some short reminder words to response for prompting people themselves to check significant facts, or also adding source annotations to question according to fact. Sometimes people could utilize technology in order to interfere with user's automatic cognitive processing instead of merely pursuing for an easy and smooth feeling. Through prompts, annotations, guide information etc., these intervention can break up automatically accepting the output from AI and makes them use more active brain power to deal with it.

In terms of education level, people need to integrate AI literacy in school. Rather than limit instruction merely to teaching how to use AI, teaching and learning activities concerning AI literacy are intended to enable students to compare information from different sources as well as identify logical gaps and potential biases in content created by AI. People must view questioning an AI as something that can be learned about and studied - not just viewed as a notion or idea.

At an individual level, the adoption of a context-based response system. In confrontations with unfamiliar territory or factual questions, people need to do several checks by themselves, other than depending on the AI. People might look up keywords or go for reliable sources. Those who have fixed views and are actively thinking them through analytically do not need more willing checks as they are thinking already [5]. What's important is not questioning everything all the time, but knowing when it is necessary to start verifying based on fact if those facts surpass what people personally know.

5.2. Future research directions

Future research can go in these directions.

First, cross-cultural research. The samples in current core empirical studies are mostly from the United States and China [1, 9, 10]. It is not a priori assumption that individualist cultures differ fundamentally from collectivist ones. Future work can examine if certain cultural orientations affect how much people trust information generated by artificial intelligence and whether they would like to verify it themselves. Cross-culture comparison on AI-trust level, frequency of verification behavior, sensitivity toward automation bias, willingness to reject recommendation made by an AI.

DOI: 10.54254/2753-7048/2026.37354

Second is longitudinal research. In existing studies most of them have tested the immediate effects after one time use of AI for example measuring if people's opinion would change in just one debate or seeing how someone learns when they write once [1, 9]. It is unknown whether a prolonged and frequent usage of AI causes gradual weakening of an individual's ability to think independently due to a lack of longitudinal data on this matter. Future studies may employ a long-term research framework to study it; examining changes in independent thinking skills, self-monitoring during learning, checking information for correctness, and using artificial intelligence at various times.

The third way is that individual difference research. Not everyone is affected by AI equally. Deng and Deng found personality dispositions influence dependency on artificial intelligence [7]. Future study can examine individual-difference variables such as disposition for reflective thinking, scepticism, trust propensity, readiness to independently verify information, etc., and it will form the basis for identifying high-risk groups.

6. Conclusion

This paper has reviewed the experimental evidence of Salvi et al. on GPT-4's persuasiveness, and discussed the psychological mechanisms behind AI persuasion and their implications for critical thinking in light of theoretical frameworks such as automation bias and cognitive laziness.

Findings show that the most noticeable persuasiveness of AI is not because it has such an "intelligent" or "objective" character itself, but due to 2 psychological tendencies which it can stir within us. The first one is cognitive laziness - people skip checking stuff themselves and just go along with what an AI says. Second is algorithm worship - tendency to transfer mental authority to AI systems by default and treat AI output as the basic factual claim basis. These two mechanisms occur both among adolescents as well as adults and could lead to less critical thinking over prolonged usage of AI. Furthermore, the statistical generating mechanism used in the creation of the AI makes it so there are no guarantees when it comes to accurate results being generated, this sort of technical shortcoming mixed up with lazy thoughts would result in individuals having difficulty in spotting false pieces of information.

In order to cope with these dangers the author has proposed 4 measures. On a national scale, establish a risk-tiered governance framework for AI; Corporational level, add verification prompts that match the task risks during interactions; On the educational side, carry out education on comparable, judgable and interpretable abilities about AI literacy; Individually speaking, it is cultivation of context dependent verification habits. Future research may further explore cross-cultural comparisons as well as longitudinal tracking and individual variations.

The main contribution of this paper is to switch the focus of discussion about AI persuasion from a technology-effectiveness perspective to one rooted in cognitive psychology, proving that over-trust's psychological origins are not due to being perfect but rather stem from individuals favoring their personal thoughts and lacking sufficient thought. This will alert people to be cautious with AI's outputs, and give some theory for teaching critical thinking education in the age of AI.

References

[1] Salvi, F., Ribeiro, M. H., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour, 9*, 1645–1653. https://doi.org/10.1038/s41562-025-02194-6

[2] Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991–1006.

DOI: 10.54254/2753-7048/2026.37354

[3] Mosier, K. L., & Skitka, L. J. (1999). Automation use and automation bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 43*(3), 344–348.
[4] Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
[5] Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning, 24*(4), 423–444.
[6] Pearson, J., Dror, I. E., Jayes, E., Whordley, G., Mason, G., & Nightingale, S. (2026). Examining human reliance on artificial intelligence in decision making. *Scientific Reports, 16*(1), Article 5345. https://doi.org/10.1038/s41598-026-34983-y
[7] Deng, Z., & Deng, Z. (2025). Becoming a cognitive miser? Antecedents and consequences of addictive ChatGPT use. *Social Science & Medicine, 383*, 118467. https://doi.org/10.1016/j.socscimed.2025.118467
[8] Shin, D. (2026). Automating epistemology: How AI reconfigures truth, authority, and verification. *AI & Society, 41*(2), 1553–1559. https://doi.org/10.1007/s00146-025-02560-y
[9] Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., . . . Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology, 56*(2), 489–530.
[10] Baldeo, S. (2026). Generative AI reliance and executive function attenuation: Behavioral evidence of cognitive offload in high-use adults. *Technology, Mind, and Behavior*. Advance online publication.