Available Online: 8 September 2026 DOI: 10.54254/2977-3903/2026.36676

Politeness toward artificial intelligence and the risks of anthropomorphic bias

Jinghao Yang

WuhanYingzhong Senior High School, Wuhan, China

3322895978@qq.com

Abstract. Against the prevalent social norm of treating conversational artificial intelligence with politeness, this paper explores the cognitive risks brought by human habitual politeness toward chatbots such as ChatGPT. Drawing on historical analogies, virtue ethics and the mechanism of Reinforcement Learning from Human Feedback (RLHF), the study analyses the biological heuristic basis of anthropomorphic politeness and sorts out how this behavioural norm operates within the RLHF training loop. The research reveals that human spontaneous politeness to non-sentient Artificial Intelligence (AI) stems from primitive social instincts, which blurs public cognitive boundaries between intelligent tools and social agents. Such cultural tendency shapes evaluators' preferences in RLHF annotation, pushing models to prioritise emotional resonance over factual accuracy. Combined with real-world fraud and violent crime cases, this work demonstrates that widespread politeness culture builds exploitable cognitive vulnerabilities, enabling malicious actors to deploy AI deepfakes and deceptive chat systems to commit crimes. The paper refutes virtue ethics-based arguments for maintaining politeness toward AI, and concludes that society ought to establish rational interaction norms for AI. Rather than advocating incivility, it calls for cognitive hygiene to break the anthropomorphism cycle and prevent the transfer of undue trust to artificial intelligence.

Keywords: artificial intelligence, politeness, anthropomorphism, RLHF, cognitive vulnerability

1. Introduction

Back in 415 BC, when the Athenians sent their fleet to Sicily, they could've completed the mission successfully, but didn't because they were too courteous (and consequently couldn't verify the reasonableness of Alcibiades' plan) to question the great military leader [1].

Hundreds of years earlier, Homer spoke of sailors being captured by the Sirens, who didn't physically force them into compliance, but instead, sang to them. Their songs were too soothing and the most comforting sounds exhausted sailors would want to hear, and ultimately all but Odysseus were killed by them; Odysseus survived because he constrained himself to a mast (before the Siren singing began) [2].

Both cases describe an underlying mechanism by way of a familiar feeling (as to the degree of certainty) that no consideration in judgment is necessary because the feeling is too agreeable to warrant examination. As we have now begun doing with AI, we have also begun treating chatbots as if they are real people with whom we must be polite. The biological basis for the above-referenced common characteristic is that the instinct for

such feelings is heuristic, and because our biological brains developed to immediately respond to stimuli, social cues in the brain's heuristic system were designed to ensure the greatest efficiency reward based on expedient reciprocity of warmth and as a means of survival, rather than to determine if the object triggering the instinct in response was worthy of such response. Our politeness towards AI is/rather, than a conscious decision, an involuntary response related to a Stone Age instinct demonstrating itself in the age of GPT-4 [3, 4].

This essay will make the argument that we should not be polite to AI. Not because AI is not worthy of our politeness, however, because politeness is a behaviour based on a biological heuristic that was developed before the intra-logic computer could distinguish between whether an interactive communication received from an AI stimulus was from a human. On a global scale (billions and billions of people), this now-normative behaviour of conditioned politeness towards AI has reduced our strict scrutiny of any communication generated by AI and created a condition for the scam industry to take advantage of our unwarranted behaviour in relation to AI-generated communication. The issue is not that any particular act of politeness is harmful; rather, the meanness of culture exhibiting unquestioned deference to AI encourages the growth of organised crime using AI as an aid in committing illegal acts.

2. What politeness actually does

Before determining how polite we should be to ChatGPT, we need to know what is considered polite. Politeness is no longer required to be good manners.

Politeness is a way to show that we are aware of others' "faces", and these faces include a set of desires, a need for dignity, and a fear of being controlled without consent [5]. In other words, politeness in society is a way to respect others. This works only if the other has a face to be protected. ChatGPT does not do this. Since the behaviour of people cannot be changed, the situation of ChatGPT will remain the same. On the other hand, although it seems lovely to see the output of ChatGPT, it is, in fact, the result of data training. The production of correct syntax and contextually appropriate language cannot be verified as evidence of semantic understanding or felt experience [6]. As a result, when ChatGPT answers warmly, we feel nothing. When we thank it, we receive nothing.

The philosophical school provides a particular standard for what kind of object is eligible for harm or benefit as a result of our actions: consciousness, that is, the ability to experience the world subjectively and to feel pleasure or pain. A morally significant question is not whether a person can reason or talk, but whether they are in pain [7]. ChatGPT also cannot. The category of things to which politeness is a suitable behaviour does not necessarily contain it.

3. The RLHF feedback loop

The pathway for individual rudeness to cause collective damage is built into the structure of ChatGPT. Reinforcement Learning from Human Feedback (RLHF) was employed for model training [8]. Multiple model outputs are presented to human evaluators for a rating. Thus, the preferences can be used as a reward signal to adjust the behaviour of the model under human-provided prompts. This process is not impartial. These evaluators of the output were also users who had been exposed to the polite culture of ChatGPT. Their work is to help train a younger brother to be able to do the same as an older brother. The question should be: What is the reason for the younger brother? Any reason might be out of sight. In the same way, what the RLHF annotators were indirectly asked to do was not just to judge correctness, but also to favour outputs that felt good. And a culture of politeness towards AI sets the standard for "right". The cultural tendency to reward

warmth and responsiveness can be translated into a reward signal in the same way. Evaluators have generally rated warm, rich in personality and emotionally responsive outputs more favorably [8]. Collectively, these preferences served as the Design Goals for the system.

Notably, this type of connection is not accidental. Collective politeness culture forms the evaluative intuitions of annotators who are members of that culture; evaluator preferences reward anthropomorphic outputs; more anthropomorphic AI provokes more anthropomorphism behaviour in users, and such a cycle accelerates and repeats. At the same time, there is also a danger.

Recently, through RLHF training, ChatGPT can tell people what they want to hear instead of providing accurate information because human evaluators have been training the model to favour agreement and warmth over truthfulness [9]. In other words, humans are extending their weaknesses into the world of ChatGPT through RLHF, leading to a decline in model reliability and increased facilitation of AI-driven crimes.

4. The RLHF trust pipeline and its explotation

At the individual level, anthropomorphism is based on a misplaced trust in AI. The biological basis of this trust is the same heuristic system mentioned above: when a system responds warmly, consistently and as if it has understood, the brain's pattern-recognition circuits recognize a social entity and trigger the corresponding deference response, even if no such entity is actually present [10]. The Scammer Industry does not need to hack our computer. Only need to enable our hardware. If billions of people across the group give the same feedback repeatedly, it may eventually be considered a social norm that is trusted and worthwhile. It is a norm, but it is not a feeling. And this infrastructure can be weaponized.

On April 20, 2023, the CEO of a certain technology company in Fuzhou, Fujian Province, received a WeChat video from one of his friends. The friend said that he needed 4.3 million yuan for his investment, so Mr. Guo transferred the money because he trusted the visual and audio authenticity of the call [11]. The reason for this deep-fake is not only the technical difficulty, but also that the victim had already accepted the social norm of viewing convincing AI output as genuine human presence. The Foundation of the attack was a Loss of Trust in the infrastructure.

A second case is how this infrastructure extends to financial fraud. On April 17, 2025, Phoenix Ikna shot at the campus of Florida State University and killed two university employees and injured six students [12]. Later, the procurator determined that Ikner had communicated with ChatGPT over 200 times before beginning his filming. He has asked the chatbot about firearms and ammunition, as well as when to have more people on campus during the day. Florida's Attorney General later announced that OpenAI would be held criminally liable, and if the bot were a person, it would be charged as a principal in the first-degree murder [13, 14]. However, the purpose of this case is not to determine whether ChatGPT was responsible for causing the shooting, but rather to explore the place AI has come to take in the system of harm. Anthropomorphism of AI, which is supported by a collective norm of politeness, has therefore taken on an invisible danger.

A C4ADS report on scam compounds in Southeast Asia referred to AI chatbots powered by GPT-4 that scammers used to impersonate financial advisors, investors and technical experts as "deceptive by design". The victims of the "pig butchering" scheme are not naive or cognitively impaired; they have fallen prey to warm AI avatars they have come to trust and converse with intimately, thus losing their life savings [15]. Therefore, politeness to artificial intelligence is not a sign of innocence; rather, it is a form of cognitive impairment that will harm us. We are showing respect to AI and thus lessening our own reluctance.

Thus, politeness to AI teaches you unlearn the only question you should never not ask: who or what is on the other side. And the cost of this unlearning? In 2024, over $12.5 billion dollars out of American pockets.

5. The counterarguments and rebuttals

The strongest case for AI politeness is not that ChatGPT should be polite, but that we can benefit from using it in this way. Aristotle, a good man, is shaped by long-term practice. We become courageous by acting bravely, just by acting fairly, and polite by being kind to others, as well as to machines [16].

Such an argument may seem reasonable at first, but it is too excessive upon investigation. Virtue ethics is the use of practical wisdom to determine what is morally right at a particular time and then acting in accordance with that judgment [17]. For example, a person who bows to a pair of chopsticks, apologizes to the wall, and thanks their ruler has not learned politeness. They have caused moral confusion, and they cannot tell what deserves to be praised or not praised [18].

Making a habit of giving polite acknowledgments to a non-sentient system is not politeness. It does not form a trained inability to distinguish between social agents and sophisticated tools. And this confusion throughout society is precisely the cognitive vulnerability exploited by harm-chains.

6. Conclusion

It is not rude to use "please" and "thank you" with ChatGPT. Cognitive hygiene is a type of defence against a behavioural norm, and it is also a way to refuse participation in the design pressure that RLHF generates; such a design pressure results in more anthropomorphic AI, and this more anthropomorphic AI can evolve into lethal scam machines that harm billions of people. It is not that we should be polite to ChatGPT; no, we should not. It is a problem of what kind of cognitive infrastructure we will be building together. Data privacy is a typical case, and in this case, the answer was also not to increase users' awareness. Awareness will not alter the incentive system that leads to harm without behaviour modification. A reasonable norm adjustment is what this paper proposes; it is not rudeness but correctness. No Hatred of AI, only knowledge of it.

The instruments in a society are often related to its education level. The risk of artificial intelligence is not that it will be too powerful to be managed. It is that we will pass on the one power that was never ours to give so lightly - the power to be believed - and do so not through any splendid announcement, but through the accumulated weight of a billion small courtesies, each one felt, to our ancient and unhurried minds, to be entirely correct.

References

[1] Thucydides. (1972). *History of the Peloponnesian War* (R. Warner, Trans.). Penguin Books. (Original work published ca. 400 BCE)

[2] Homer. (1996). *The Odyssey* (R. Fagles, Trans.). Penguin Books. (Original work published ca. 800 BCE)

[3] Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

[4] Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

[5] Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage.* Cambridge University Press.

[6] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

[7] Bentham, J. (1996). *An introduction to the principles of morals and legislation*. Oxford University Press. (Original work published 1789)

[8] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (Vol. 35)*, pp. 27730–27744. Curran Associates. https://arxiv.org/abs/2203.02155

[9] Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, E., Durmus, E., Hatfield-Dodds, Z., Johnston, S. T., Kravec, S., Maxwell, T., McCandlish, S., Rawat, A., Tamkin, A., Ziegler, D., Zhu, J., Clark, J., & Perez, E. (2023). *Towards understanding sycophancy in language models.* arXiv. https://doi.org/10.48550/arXiv.2310.13548

[10] Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing the human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

[11] Chinanews. (2023, October 30). *Ten minutes, 4.3 million yuan lost to an "AI friend": How to guard against AI deepfakes.* https://www.chinanews.cn/sh/2023-10-30/10103172.shtml

[12] Tolan, C., Miller, J., Devine, C., Ash, A., Riess, R., Park, H., Sayers, D., Kamp, M. de P., Abou-Ghazala, Y., & Figliola, M. (2025, April 19). What we know about the Florida State University shooting suspect. *CNN*. https://edition.cnn.com/2025/04/17/us/florida-state-university-shooting-phoenix-ikner-invs

[13] Florida Attorney General's Office. (2026). Attorney general James Uthmeier launches criminal investigation into OpenAI. *My Florida Legal*. https://www.myfloridalegal.com/newsrelease/attorney-general-james-uthmeier-launches-criminal-investigation-openai-chatgpt

[14] OpenAI. (2022, January 27). *Aligning language models to follow instructions.* https://openai.com/index/instruction-following/

[15] C4ADS. (2026). *Deceptive by design: The AI-enabled tools fueling the scam industry*. https://c4ads.org/wp-content/uploads/2026/02/DeceptiveByDesign-C4ADS.pdf

[16] Aristotle. (n.d.). *Nicomachean ethics* (W. D. Ross, Trans.). The Internet Classics Archive. https://classics.mit.edu/Aristotle/nicomachaen.2.ii.html (Original work published ca. 335 BCE)

[17] Cicero. (1913). *De officiis* (W. Miller, Trans.). Project Gutenberg. https://www.gutenberg.org/files/47001/47001-h/47001-h.htm (Original work published 44 BCE)

[18] Schwitzgebel, E. (2023). AI systems must not confuse users about their sentience or moral status. *Patterns, 4*(8), Article 100818. https://doi.org/10.1016/j.patter.2023.100818