Proceedings of the 8th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2026.37156

Politeness to ChatGPT Reconsidered: Structural Injustice and Respect for Rational Agents

Xiaokai Wen

Christ Episcopal School, Covington, USA
xiaokaiwen537@gmail.com

Abstract. The widespread interaction between humans and large language models has sparked normative debates regarding user politeness toward artificial intelligence. This paper explores the ethical rationality of maintaining polite linguistic norms in human-LLM interaction, arguing against the normative obligation of being polite to artificial intelligence. It analyzes the social risks triggered by normalized AI politeness, demonstrating that such social norms drive technology enterprises to form fixed politeness-conditioned design under the constraint of AI value alignment mechanisms. This design centered on English cultural logic leads to differential algorithmic treatment of users from diverse cultural, linguistic and physical ability backgrounds, resulting in structural social injustice. From the perspective of Kantian moral philosophy and capability ethics, this study verifies that the algorithmic differential treatment essentially disrespects human rational agency and erodes individual autonomous capacities. The research concludes that practicing politeness toward large language models is ethically inappropriate due to its inherent injustice, and such interactive norms should not be universally advocated.

Keywords: AI politeness, algorithmic injustice, value alignment, rational agency, human-AI interaction

1. Introduction

Addressing ChatGPT with rudeness or incivility causes the output to consume an additional 14.4 tokens, even when the substantive content of the response remains nearly identical [1]. Concurrently, a 2025 study from Penn State University demonstrated that, under highly impolite prompts, GPT-4o achieved an accuracy of 84.8%, whereas highly polite prompts yielded only 80.8% [2]. This raises a moral dilemma: Should we base our guidelines on whether to treat AI with courtesy on resource considerations or on accuracy? Yet such findings describe only how the model behaves; they cannot settle how we ought to treat it.

A fundamentally normative question needs to be raised: Ought individuals to show politeness toward large language models, with ChatGPT foremost among them? What does politeness mean? Should society impose such a normative requirement? My conclusion is that we should not be polite to LLMs. This essay defends this conclusion in two steps. First, I will argue that being polite to AI gives rise to injustice, because the normalization of politeness toward AI inevitably leads companies to adopt politeness-conditioned design, which in turn produces differential treatment of users.

© 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: 10.54254/2755-2721/2026.37156

Second, I will argue that we ought not to do what gives rise to injustice, on the grounds that injustice is a failure to respect persons as rational agents. From these two premises, I conclude that individuals ought not to be polite to AI.

2. Injustice of AI politeness

Although being polite to AI may appear to be a matter of user choice, it in fact has some broader social implications. When determining whether being polite to AI should be adopted, the key issue is whether those social implications are ethically permissible. If they are not acceptable, then being polite to AI should not be adopted. In this section, I will argue that being polite to AI may lead to injustice, because it necessarily compels companies to engage in politeness-conditioned design, thereby producing culturally differentiated responses.

A user's attitude toward AI affects not only the performance of the AI itself but also its design, and such design is mediated by the normative requirements of value alignment. Value alignment requires the AI system to remain consistent with the value system of human society. Frames it as preventing machines from acquiring objectives imperfectly aligned with our own [3]. Moreover, value alignment encompasses both abstract ethical principles (such as non-maleficence and respect for autonomy) and concrete social norms (such as interactional etiquette and accepted patterns of conduct). The goal of value alignment is therefore to map the norms of interpersonal interaction onto the human-machine relationship, so that as the former evolves, the latter is guided to evolve in step.

Politeness, too, is a key requirement of value alignment, demanded both at the level of principle and at the level of practice; companies therefore engage in politeness-conditioned design. For example, OpenAI released its Model Spec in May 2024 and has continued to update it [4], explicitly listing "encourage fairness and kindness" among ChatGPT's default behaviors which is an instance in which developers actively encode the alignment of AI behavior with a social norm into binding behavioral constraints on the model. Accordingly, when people consider being polite to AI as a social norm, value alignment leads developers to hold that AI ought to conform to the norms of human society, and they will therefore actively design polite responses. Moreover, on the side of public governance, AI be aligned with human values [5], which indicates that value alignment is the prevailing paradigm of international AI ethical governance. Consequently, once politeness becomes a social norm, value alignment obliges companies to proactively adopt politeness-conditioned design.

Furthermore, once this design is in place, the AI will inevitably respond differently to individuals from different cultures and different backgrounds. To speak is not merely to state a fact, but also to perform a certain act (such as a request, a command, or a promise) [6]. Politeness is a linguistic function deployed for the purpose of achieving an end.

Successfully achieving that end requires speaker and hearer to share a cultural consensus. For example, in English, moderate politeness combined with directness is considered appropriate. But in Japanese, the honorific system is intricate, and only an elevated degree of politeness is considered appropriate. However, AI design tends to take English as its primary cultural consensus and ignoring the cultural variability of other languages. The effect of prompt politeness on large-language-model performance varies considerably across English, Chinese, and Japanese [7]. Research on LLM training data confirms that most large language models are predominantly English-centric [8]. Yet AI companies routinely use English as the baseline for designing their politeness templates, with the result that politeness- conditioned design ultimately delivers differential responses to users from different cultural backgrounds.

DOI: 10.54254/2755-2721/2026.37156

This differential response is in itself a form of injustice. Justice requires that all social positions and offices be open to everyone under conditions of fair equality of opportunity, and the principle of fair equality of opportunity demands that no individual in society obtains different opportunities on account of morally arbitrary attributes [9]. Morally arbitrary attributes refer specifically to characteristics determined not by personal choice or merit but by contingent factors (such as birth, race, gender, cultural background, and natural endowment). This principle therefore requires that institutions remain neutral with respect to morally arbitrary attributes. Consider, for example, whether disabled or neurodivergent users, whose attributes are morally arbitrary, ought to be polite to AI. For ordinary users, being polite to AI may merely be a matter of habit, but for some disabled or neurodivergent users, such as those with motor impairments, exchanging politeness with AI may demand far greater effort and time.

Accordingly, the differential treatment provided by AI is an injustice grounded in morally arbitrary attributes, because it causes the same request to receive different treatment on the basis not of the user's actual needs or abilities, but of arbitrary attributes such as language, culture, and politeness style, which are irrelevant to its legitimacy and therefore constitute a form of injustice.

Some may argue that even disabled or neurodivergent users who possess morally arbitrary attributes ought still to be polite to AI. This is because politeness is a virtue, and even neurodivergent users or speakers of minority languages ought to practice this virtue, and should not treat AI carelessly. This view appears to have some merit, but it overlooks the fact that the core of virtue's requirement is situational and proper judgment rather than formal compliance. Here we should distinguish the ability of politeness from the character of politeness: the former is merely knowing how to follow a rule, while the latter is being polite for a reason. In Book VI of the *Nicomachean Ethics*, distinguishes between two kinds of knowledge: purely rule-based episteme and situationally attuned phronesis, and it is the latter that truly constitutes practical virtue, requiring the agent to respond appropriately to the particular situation at hand. An aphasic patient who conveys gratitude through eye contact and an articulate person who uses honorifics stand equal in virtue, for each has fulfilled the same inner character of "treating others with courtesy" [10].

However, once the social norm of being polite to AI is concretized in politeness- conditioned design, it strips away the dimension of phronesis, reducing politeness to a single external template imposed on everyone—Parkinson's patients with input difficulties, autistic users who struggle to produce social pleasantries, or those reliant on eye-tracking keyboards [11]. Even users inwardly committed to courtesy are judged insufficiently polite and receive worse responses; situational judgment gives way to a uniform standard, and formal compliance masks differences in capability. Worse still, by cloaking this standard in the language of virtue, the system re-describes the disadvantages of neurodivergent and minority-language users as deficiencies in virtue: they shift from victims of an unjust system to people blamed as not polite enough, so that structural injustice is transferred onto the very individuals it harms.

3. We ought not to do what gives rise to injustice

However, to show that being polite to AI gives rise to injustice does not establish that we ought not to be polite to AI. To close this gap, in this section, I will argue that people ought not to do what gives rise to injustice, because injustice is disrespect toward rational agents, and we ought not to do what disrespects rational agents.

The essence of justice lies in the differential dispossession that institutions impose upon rational agents. Differential dispossession refers to institutions stripping the individual, as a rational agent, of the right, to explain and defend themselves. Injustice is disrespect toward the individual's standing

DOI: 10.54254/2755-2721/2026.37156

as a rational agent, including the loss of the right to defend oneself. In other words, the essence of injustice is a form of disrespect. A person is an end in itself and never a mere means [12]. That is to say, to respect someone is to grant them the equal standing of a rational agent; to use someone merely as a means denies that standing.

For example, COMPAS is a machine-learning recidivism risk assessment tool used in the United States' judicial system. Some reports demonstrate that COMPAS assigned significantly higher risk scores to black than to white defendants, while operating as an unexaminable black box [13]. A defendant could neither learn the basis of the score nor challenge it, and was compressed into a "probability of reoffending," deprived of all space, as a rational agent, to demand explanation, raise objections, or account for their circumstances. COMPAS treats the defendant as a means of judicial efficiency, demoting a subject capable of reasoning and self-account into a data point that passively receives a verdict. Politeness-conditioned AI design follows the same logic, reducing users who fall outside dominant cultural norms to deficient inputs rather than treating them as rational agents deserving equal consideration.

It must next be shown that we ought not to do what disrespects rational agents. A rational agent possesses the capacities for autonomy, critical reflection, and social participation. Disrespect, accordingly, is the obstruction, denial, or appropriation of these capacities, denying the agent as an active agent of capacities.

Take an example where AI assists for clinical decisions. The ACCEPT trial in Poland found that after an AI-assisted polyp detection system was introduced in late 2021, the adenoma detection rate of endoscopists working without AI fell from 28.4% to 22.4%, with AI exposure confirmed as an independent factor [14]. Sustained reliance on AI thus intercepts physicians' own capacity for judgment and supplants its exercise. Every agent ought to have the opportunity to develop their own capacities, and technology that turns the person into a passive recipient erodes the capacities of agency and subjectivity. To erode these capacities is to appropriate what defines a rational agent, and such appropriation is precisely the disrespect this section sets out to establish [15].

4. Conclusion

To sum up, through value alignment, politeness-conditioned design produces injustice toward users from different backgrounds, while injustice is Kant's disrespect toward rational agent. We ought not to do because it appropriates the agent's capacities and constitutes disrespect. Therefore, we ought not to be polite to AI.

If future AI design were genuinely able to escape the English-baseline template of politeness, to achieve cross-linguistic and cross-cultural neutral alignment, and to ensure that politeness no longer translates into differential treatment, then the argument of this paper would need to be reexamined. Until that day arrives, however, politeness ought not to be entrusted to a system that would produce injustice on account of it.

References

[1] Alavi, S., Nozari, S., & Luangrath, A. (2025). *Cost transparency of enterprise AI adoption*. arXiv. https://arxiv.org/abs/2511.11761

[2] Dobariya, O., & Kumar, A. (2025). *Mind your tone: Investigating how prompt politeness affects LLM accuracy*. arXiv. https://arxiv.org/abs/2510.04950

[3] Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

[4] OpenAI. (2024, May 7). *Introducing the Model Spec*. https://openai.com/index/introducing-the-model-spec

DOI: 10.54254/2755-2721/2026.37156

[5] UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence
[6] Austin, J. L. (1975). *How to do things with words* (2nd ed.). Harvard University Press.
[7] Yin, Z., Wang, H., Horio, K., Kawahara, D., & Sekine, S. (2024). *Should we respect LLMs? A cross-lingual study on the influence of prompt politeness on LLM performance*. arXiv. https://arxiv.org/abs/2402.14531
[8] Wang, Y., Min, S., Xu, K., et al. (2023). *Turning English-centric LLMs into polyglots: How much multilinguality is needed?* arXiv. https://arxiv.org/abs/2312.12683
[9] Rawls, J. (1999). *A theory of justice* (Rev. ed.). Harvard University Press.
[10] Aristotle. (2000). *Nicomachean ethics* (R. Crisp, Trans.). Cambridge University Press. https://doi.org/10.1017/CBO9780511802058
[11] Kalaganis, F. P., Chatzilari, E., Nikolopoulos, S., Kompatsiaris, I., & Laskaris, N. A. (2018). An error-aware gaze-based keyboard by means of a hybrid BCI system. *Scientific Reports, 8*(1), 13176.
[12] Kant, I. (2012). *Groundwork of the metaphysics of morals* (M. Gregor & J. Timmermann, Eds. & Trans.). Cambridge University Press. https://doi.org/10.1017/CBO9781139032490
[13] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[14] Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., Blom, J., Buszkiewicz, M., Halvorsen, N., Hassan, C., Romańczyk, T., Holme, Ø., Jarus, K., Fielding, S., Kunar, M., Pellise, M., Pilonis, N., Kamiński, M. F., Kalager, M., Bretthauer, M., & Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet Gastroenterology & Hepatology, 10*(10), 896-903. https://doi.org/10.1016/S2468-1253(25)00133-5
[15] Nussbaum, M. C. (2011). *Creating capabilities: The human development approach*. Belknap Press of Harvard University Press.