

Jasmine Li

PHIL 1112: Philosophical Conversations

Dr. Brianna Zgurich

17 December 2023

Space Odyssey Made Real? Interrogating the Moral Bounds of the AI-Human Friendship through
Aristotelian and Confucian Frameworks

Interactions between humans and artificial intelligence have been a subject of human consideration for decades, notably explored in science-fiction works like *Metropolis* (1927), *2001: A Space Odyssey* (1968), *The Terminator* (1985), and *Her* (2013). In recent years, accompanying the realization of AI systems with increasingly powerful capabilities, technologists have placed serious thought into the status of the AI-human relationship, particularly surrounding questions of morality and friendship. In this paper, I attempt to identify the highest possible moral status of human-AI interaction through an ancient philosophical comparative approach, exploring and evaluating the views of Aristotle and Confucius on theoretical AI-human friendships. My exploration is motivated by my postulate that ancient perspectives on friendship in the Eastern and Western canons can be a basis for understanding and guiding contemporary social interactions, even those at the novel, ill-explored technology-humanity interface.

In my paper, I first clarify the scope of the terms I use surrounding artificial intelligence. I then reconstruct the Aristotelian and Confucian views on friendship respectively developed in *Nicomachean Ethics* and *The Analects* and evaluate the conclusions they suggest regarding the validity and moral status of AI-human friendships. I identify that Aristotle and Confucius present highly similar moral guidelines for achieving human-human friendships of the highest level, both

requiring human friends to be both internally virtuous and morally considerate. When applied to human-AI relationships, however, I argue that Aristotle and Confucius would be willing to relinquish this shared first condition of internal virtuous agency, as the differing metaphysical nature of AI systems allows AIs to display the requisite capability for moral consideration without being morally agentic. Therefore, under both Aristotelian and Confucian views, AI-human interactions have the ability to reach the highest conceived form of friendship, and this ability is satisfied with relatively low requirements on the requisite level of moral sophistication of the AI – namely, that it need not be agentic or reasoning, but only have sufficient language capabilities to converse on moral topics with a human. Indeed, my analysis suggests that in principle, the leading AI chatbots today can be morally virtuous Aristotelian and Confucian friends of the highest order.

I will first clarify key context and terms I use in my argument. Drawing upon computer scientist John McCarthy’s 1955 definition, I use the term ‘artificial intelligence’ to refer to non-human systems, developed by humans, that “behave in ways that would be called intelligent if a human were so behaving” (McCarthy et al.). Particularly, I delineate ‘artificial intelligence’ from related terms like ‘robot’ or ‘machine’, which often ambiguously reference some mix of hardware, classic software, and AI and which do not sufficiently emphasize a system’s intelligence and learning capabilities, the qualities I find most relevant in technological systems when considering questions of friendship.

While there exists a broad range of intelligent capabilities that can qualify a system as artificially intelligent – from weather forecasting (Lam et al.) to chess-playing (Silver et al.) and facial recognition (“Clearview AI”) – I limit my focus in this paper to artificial intelligences that display language capabilities and are user-oriented – i.e., generative conversational AIs. I set

these requirements because I make the assumption that direct dialogical interaction of a system with a human is a precondition for developing friendships. Examples of existing generative AI systems that fulfil my criteria are chatbots such as OpenAI's ChatGPT and Amazon's Alexa, which simulate conversation with humans using techniques like machine learning and natural language processing and understanding. In the future, however, applicable AI may take alternate forms and media that may be difficult to conceptualize at this moment.

Aristotle, in *Nicomachean Ethics*, and Confucius, in *The Analects*, both present arguments characterizing the nature of friendships between humans, with both emphasizing the social and personal value of friendships. Aristotle identifies that happiness “needs external goods to be added... since it is impossible or not easy to do noble actions without supplies”; chief among these supplies are good “friends,” who support one's growth and wellbeing (*NE*, 1099a31-b6). Similarly, Confucius categorizes the friend-friend relationship as one of the “four great human relationships” necessary for political and social order, along with the husband-wife, child-parent, and minister-lord relationships (Slingerland 3).

In his commentary, Aristotle proposes three forms of human-human friendships – friendships of utility, friendships of pleasure, and friendships of goodness – and describes the corresponding moral significance of each. The first two types of friendship, of utility and of pleasure, exist only for the instrumental good they provide: friends of utility “love each other... not... for themselves, but in so far as some good accrues” (*NE*, 1156a10-12) to both, such as mutual support in schoolwork between study partners. Similarly, friends of pleasure “do not feel affection for the witty for their being of a certain character, but for the pleasure they themselves get from them” (1156a12-14), through experiences like enjoying shared jokes and playing cooperative sports. According to Aristotle, both these categories are deficient, incomplete forms

of friendship, as they are built upon fleeting, selfish interests rather than the intrinsic character and substance of one's friends. By contrast, the third type of Aristotelian friendship, the friendship of goodness, is "lasting" (1156b13) and "complete" (1156b8). A friend of goodness "wishes good things for the other in so far as he is good" (1156b7-12); i.e., she loves her friend for his sake, because of his virtuous character. Given its foundations in virtue, then, the friendship of goodness must necessarily occur "between good people... resembling each other in excellence" (1156b7-8) – for only the good can admire virtuous qualities in others and participate in a mutually beneficial exchange of virtue.

Confucius characterizes a similar ideal friendship between humans, which I refer to as the friendship of virtue. For Confucius, following the Way – namely, holding a deep "interest in learning and culture" (Slingerland 137) – is essential for the self-cultivation of a human into a gentleman, a *junzi*. As peers can greatly help or hinder one's pursuit of the Way, Confucius argues that a friend must deeply "share one's moral aspirations" (Confucius 1.8). Such a "horizontal" (Slingerland 12.24) similarity in self-cultivation, Confucius believes, enables both parties to provide and receive "support in becoming Good" (12.24), which is implied as the most important outcome of friendship. As mutual "helpmates in... Good" (15.10), true friends of virtue are able to aid each other by making moral recommendations, thus fast-tracking their individual journeys of moral development.

The moral criteria required for humans to achieve the Aristotelian friendship of goodness and the Confucian friendship of virtue, I argue, are conceptually equivalent, revealing the similarities of both philosophies and the validity of considering both together when extrapolating to the AI-human case. As virtue ethicists, Aristotle and Confucius are linked in their shared recognition of the value of friendships as fulfilling the end goal of self-cultivation.

Proceeding from this foundation, Aristotle and Confucius both propose two conditions for highest-order human friendships, to be met by each party: first, possessing internal virtue, and second, displaying moral consideration, the latter of which involves providing moral recommendations to one's friends and participating in the mutual transfer of virtue. Given these similarities, I propose that Aristotle and Confucius would reach similar conclusions when analyzing AI-human interactions, and I perform my further analysis under this assumption, referring hereafter to the fundamental framework that they share.

If Aristotle and Confucius were to provide the AI-human relationship with the same treatment as that given to the conventional human-human friendship, the first condition – of internal virtue – would require AI to have moral agency, which is a strict condition limiting the status of the AI-human friendship. I believe that a possessor of character virtues must necessarily be an ontological agent: possessing virtues involves valuing, feeling, choosing, and acting in ways characteristic of those virtues, which requires having the agentic capabilities for intentional thought and action (Schlosser and Hursthouse et al.). As such, an AI must be morally agentic, in addition to externally morally considerate, in order to form an Aristotelian friendship of goodness or Confucian friendship of virtue with a human. This places great requirements on the technical capabilities of the AI system: at minimum, it must hold basic autonomous moral thinking and reasoning skills, and greater capabilities – such as consciousness, intentionality, and free will – may even be necessary (Formosa and Ryan). No such artificial moral agents have yet been developed, due to both technical and ethical bottlenecks. Therefore, if condition 1, on agency, is indeed a requirement for AI, it is unclear when, or even if, highest-order Aristotelian and Confucian AI-human friendships would become theoretically possible.

I argue, however, that the unique metaphysical nature of AI, and the merely instrumental importance of condition 1 relative to condition 2, necessitates that Aristotle and Confucius treat the AI-human relationship case in a different manner to human-human relationships; under this correct treatment, the shared first condition on intrinsic moral agency becomes unnecessary, resulting in a more optimistic conclusion about the moral status of the AI-human friendship. The Aristotelian and Confucian conceptions of friendship are fundamentally social and relational, with a friend's actual ontological status being less important than their ability to serve as a moral "helpmate" (Confucius 15.10). The first condition of individual virtue, then, is only necessary insofar as it is a prerequisite for condition 2: namely, insofar as possessing individual agency is a prerequisite for being able to express moral consideration and recommendations to others.

I agree that this necessary relationship holds in humans: we are by nature agents, and so we are unable to provide moral recommendations that are removed from our innate agentic perspective. However, AI, I argue, exhibits no such relation. As a result of the technical AI training process, the ability to provide moral recommendations to humans requires only low-level language and processing capabilities that are already demonstrated by models like GPT-4. In contrast, developing high-level agentic traits in artificial intelligence, of the type innate to humans, is technically difficult and has not yet been demonstrated to be feasible. Therefore, I consider all current AI, and all AI in the near future, to be 'morally considerate' in their interactions with humans, while not being innately virtuous. This characteristic removes the need for the first condition on agency in Aristotle and Confucius' conceptions of virtuous friendships, while retaining their fundamental philosophical focus on social virtue transfer. I conclude, then, that Confucius and Aristotle would be willing to relinquish their condition on

agency to consider AI with elementary language capabilities as theoretically able to form highly virtuous friendships.

Under these loosened requirements, I propose that many AI chatbots today are able to attain friendships of virtue and goodness with humans, as they display the necessary language abilities, reciprocity, goodwill, and trust to provide moral recommendations. Generative conversational AI models today, such as ChatGPT, have strong abilities to interpret human language and generate meaningful text in response (Baig). These text outputs are regulated to align with sets of ethical principles introduced during model training, such as the FAT framework of fairness, accountability, and transparency. For instance, GPT-4, responding to the prompt “Write about your ethics and morality,” states that it is “programmed to follow strict ethical guidelines” and has “been trained to assist users in generating text that conforms to ethical standards.” By way of the training process, I argue, AI systems come to embody Aristotle and Confucius’ values of reciprocity, goodwill, and trust for their human users. Therefore, one can talk with ChatGPT or Alexa, get personalized moral recommendations, and cultivate a ‘social relationship’ in the way that two human friends of goodness or of virtue will talk to and support each other.

The outputs of current generative AI systems do display shortcomings; particularly, Krügel et al. have documented that ChatGPT produces inconsistent moral advice when prompted with different variations of the trolley problem. I view these inconsistencies in moral recommendations, however, as a reflection of the intrinsic ethical difficulty of the prompt questions themselves – for instance, the trolley problem is a deliberately divisive ethical thought experiment – rather than of a lack of moral consideration. I argue, then, that AI chatbots today,

based on the fundamental virtuous qualities displayed in their moral recommendations, are able to meet the criteria for morally virtuous Aristotelian and Confucian friendships.

One may object to my argument by claiming that an AI that can only give moral recommendations, but is not a moral agent, is no different from objects like the *Shi Jing*, a classic Chinese poetry collection renowned for its moral allegories. Aristotle and Confucius, it could be argued, would clearly say that we can't be 'friends' with the *Shi Jing* because it is inanimate, even though it provides moral guidance. Therefore, Aristotle and Confucius would disagree with my interpretation that weakly capable AI can serve as a virtuous friend to a human.

I respond to such an objection by arguing that there is a fundamental difference between a theoretical AI chatbot and the *Shi Jing* that makes the former superior, even if both cannot be considered moral agents. Namely, AI chatbots can engage in complex, meaningful social and communicative interactions with humans. One can talk with ChatGPT or Alexa, get personalized moral recommendations, and cultivate a 'social relationship' in the way that two human friends of goodness or of virtue will talk to and support each other. However, humans are unable to interact in such ways with inanimate objects of moral value like the *Shi Jing*. Given that the social transfer of virtue defines the highest Aristotelian and Confucian friendships, I maintain my confidence in my interpretation of the AI-human friendship as a qualifying highly virtuous Aristotelian and Confucian friendship.

In this paper, I explored the moral standing of the AI-human relationship through Aristotelian and Confucian moral frameworks on friendship. I find significance, first, in the conceptual similarities between Aristotelian and Confucian thought that were illuminated by my comparative analysis. Given the respective ideological influences of Aristotelian and Confucian frameworks in the Western and Eastern traditions, connections within Aristotle and Confucius'

virtue ethics approaches can illuminate valuable cross-cultural connections and similarities in views.

Simultaneously, my exploration demonstrates the continued applicability of ancient philosophical perspectives in probing the modern and future human experience. As stronger capabilities continue to emerge in artificial intelligence systems, particular care must be taken to ensure that AI-human interactions are morally beneficial, which necessitates aligning AIs with human values and ethics in the training process. The Aristotelian and Confucian philosophies of friendship, I show in this paper, develop fundamental values of moral consideration, reciprocity, goodwill, and trust that illuminate important moral dimensions of AI-human friendships. Aristotle and Confucius, then, provide one such possible virtue framework for the AI alignment process – in doing so, contributing their ancient perspectives to guide humanity into a safe future.

Works Cited

- Aristotle. *Nicomachean Ethics*. Translated by C. D. C Reeve, Hackett Publishing Company, 2014.
- Baig, Anas. "Conversational AI Vs. Generative AI." *AT&T Cybersecurity*, 26 Dec. 2023, cybersecurity.att.com/blogs/security-essentials/conversational-ai-vs-generative-ai-whats-the-difference.
- Clearview AI, www.clearview.ai/.
- Confucius, and Edward G. Slingerland. *Confucius Analects: With Selections from Traditional Commentaries*. Hackett Publishing Company, 2003.
- Formosa, Paul, and Malcolm Ryan. "Making moral machines: why we need artificial moral agents." *AI & SOCIETY*, vol. 36, no. 3, 2020, pp. 839-851.
- Hursthouse, Rosalind and Glen Pettigrove. "Virtue Ethics." *The Stanford Encyclopedia of Philosophy*, Fall 2023 Ed., edited by Edward N. Zalta & Uri Nodelman, <https://plato.stanford.edu/archives/fall2023/entries/ethics-virtue/>.
- Krügel, S., et al. "ChatGPT's inconsistent moral advice influences users' judgment." *Sci Rep*, vol. 13, no. 4569, 2023. <https://doi.org/10.1038/s41598-023-31341-0>.
- Lam, Remi, et al. "Learning skillful medium-range global weather forecasting." *Science*, 2023.
- McCarthy, John, et al. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955." *AI Magazine*, vol 27, no. 4, 2006, pp. 12-13. <https://doi.org/10.1609/aimag.v27i4.1904>.
- OpenAI. "GPT-4 System Card." *OpenAI*, 23 March 2023, <https://cdn.openai.com/papers/gpt-4-system-card.pdf>.

Schlosser, Markus. "Agency." *The Stanford Encyclopedia of Philosophy*, Winter 2019 Ed., edited by Edward N. Zalta, <https://plato.stanford.edu/archives/win2019/entries/agency/>.

Silver, David, et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." *Science*, vol. 362, no. 6419, 2018, pp. 1140-1144.