Interpreting Ordinary Uses of Psychological and Moral Terms in the AI

Hyungrae Noh, Sunchon National University (South Korea)
International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: Philosophy of artificial intelligence; ordinary language; semantics; agency; moral patiency; emotions; social habits

Extended Abstract

Contemporary laypersons (henceforth, 'we') show emotional responses to AI robots. We also sometimes explain AI robots' behavior using psychological terms, suggesting that our emotional responses to them are not solely related to how we feel about them, but also to how we evaluate their behavior. Moreover, we treat AI robots as moral patients. In short, in the ordinary uses of language, we use psychological and moral terms when we evaluate the behavior of AI robots. Yet, intuitively, we would expect proper referential extensions of psychological and moral terms to exclude artifacts, raising the question of whether such referential shifts from the human domain to the AI domain follow semantic changes. What do we mean when we agree with sentences like 'AI robots believe something,' 'AI robots feel emotions,' or 'AI robots should not be harmed?' This paper discusses whether the concepts we express by psychological and moral terms in reference to AI robots are similar to those of human cases.

It is standard to hold that the meaning of a term is the concept it expresses. The conceptual structure of a term, however, can vary depending on context. Consider the concept 'moral patient,' meaning that, other things being equal, moral patients have their own rights not to be harmed. Suppose that this concept is expressed if moral terms express their literal meaning in the relevant contexts. Kantian ethics holds that, in mistreating a dog, I do not violate any obligations I owe to the dog, but I violate a duty I owe to myself, which is to cultivate morally good dispositions. If I am a Kantian, there is a genuine semantic difference between my saying that 'you should not harm children' and 'you should not harm dogs,' such that the concept of moral patiency is literally expressed only in the former. In the latter, I am expressing a technical
moral concept that you have an indirect obligation in regard to dogs that you owe to yourself. On the other hand, for Utilitarians, any sentient beings are moral patients. So, concerning the ordinary uses of the term form 'should not harm,' whether nonhuman animals are excluded or included in the proper moral extension is determined by the concept that a speaker expresses. In sum, regarding the term form 'should not harm' in the following sentences 'you should not harm children' and 'you should not harm dogs,' the referential shift follows a semantic change for Kantians, but it preserves semantics for Utilitarians.

Terms do not have fixed meaning. For instance, the term 'marriage' traditionally denoted a relation of opposite-sex couples, but now extends to same-sex couples. Although artifacts are intuitively excluded from the proper extension of psychological and moral terms, we nevertheless might literally be considering them to be psychological or moral beings when we

apply the relevant terms to AI robots. And, determining whether we really consider them as so requires an analysis of our commonsensical understanding the conceptual structures of the terms (i.e., the technical view), and a psychological analysis of the way that we extend the terms to AI robots (i.e., the habit and emotion views).

It is worth stressing that I distinguish between two approaches to referential shifts, namely a justificational approach and an interpretational approach, and that this paper primarily concerns the latter. A justificational approach attempts to find the proper domain of concepts by presupposing appropriate rules for the concepts' use. Consider the example where cognitive neuroscientists using the psychological term 'decide' in reference to the brain's (or parts of the brain's) information processing. Bennett and Hacker (2022) argue that the ascription of psychological attribute 'decide' to the brain (or parts of the brain) is misconception because it makes no sense to ascribe such attribute to anything less than the human (or the animal) as a whole. According to them, nonsense is generated when an expression is used contrary to the rules for its use, and the rules can be elicited from its standard employment and received explanations of its meaning. They claim that the rules for psychological concepts' use include only of a human being and what resembles (behaves like) a living human being can one say it has mental properties. In short, according to Bennett and Hacker, the concept's use in the cognitive neuroscience context is not justified, thus the referential shift from the human domain to the brain domain follows a semantic change. On the other hand, Figdor (2018) attempts to make sense of the referential shift. She claims that it is unclear what properties and relations we denote by psychological concepts; and, epistemic justification of our pre-theoretic division between the human and nonhuman domains is missing. She claims that the same scientifically discovered structures across the relevant human and nonhuman domains can account for the rules for psychological concepts' use. She argues that the referential shift implies that philosophers should consider expanding the proper extension of the term 'decide' accordingly because the way in which the brain (or parts of the brain) 'decides' is in fact explained by the mathematical model originally introduced to explain decision-making of human beings. In a nutshell, according to Figdor, cognitive neuroscientists are justified in using the psychological term 'decide,' therefore the referential shift preserves semantics.

An interpretational approach attempts to answer the question of what we (i.e., laypersons) really mean by psychological and moral terms used in unexpected domains. Consider again the term form 'should not harm' used in the human and nonhuman-animal domains respectively: Kantians intend to express different concepts in each context. One way of giving an interpretational approach to the uses of the term is to discuss whether we are Kantians. Similarly,
in spite of using the same psychological and moral term forms in the human and AI domains, we might intend to express different concepts in each context. The objective of this paper is to determine whether *the unexpected referential shift* from the human domain to the AI domain preserves semantics. In other words, this paper discusses whether we use the relevant terms in the AI domain 'literally' in the sense that the terms express the concepts that they typically express in the human domain.

My interpretational approach is a meta-analysis of experimental results in the study of the unexpected referential shift. Consider Huebner's (2010) experiment. He analyzed laypersons' intuitions with respect to the ascription of belief, pain, and happiness to robots. In his work, Huebner seems to assume that participants used the relevant terms with their literal meaning in their reports. For instance, if a participant strongly agrees with the statement 'a robot feels

pain if it is damaged in some way,' she is assumed to be expressing the concept of feeling pain. In this paper, I discuss tentatively plausible alternative nonliteral interpretations of this report. For instance, consider the following interpretation: 'I do not literally mean that the robot feels pain, but what I really intend to say is that if the robot is damaged, it's program will make it behave as if it feels pain.' According to this interpretation, the participant expressed a technical sense of 'designed-function' in the report, therefore semantics is not preserved.

To sum up, an interpretational approach finds plausible interpretations of the unexpected referential shift in the ordinary uses of language. A literal interpretation denies any significant semantic change: when psychological and moral terms are used in reference to AI robots, we express concepts similar to those typically expressed when the terms are used in the human domain. A nonliteral interpretation accounts for semantic changes, according to which, concerning the relevant terms used in the AI domain, we express concepts that are quite different from the literal meaning. Recall that interpreting is one thing, and justifying is another. Consider again the concept of decide. A justificational approach discusses whether AI robots should be included in the proper domain of this concept. I, however, remain neutral whether AI robots are indeed qualified psychological or moral entities in this paper. In other words, I do not attempt to show that robots make decisions just like the way that we make decisions (despite my claim that the unexpected referential shift concerning agency-terms preserves semantics).

Nevertheless, my literal interpretation can be further argued to support a theory of justificational approach. Consider Coeckelbergh's (2011) version of a justificational approach, namely the relational approach to human-robot relations. He claims that "the appearance of robots in human consciousness is mediated by language: how we use words interprets and co shapes our relation to others—human others or artificial others" (2011: 62). For example, the moral status of AI robots is constructed and grows on the basis of the relations we have with them as epistemic subjects, and thus our linguistic practices capture the cutting edge of the moral status of AI robots. To put this otherwise, if it turns out that there is no plausible nonliteral interpretation, proponents of the relational approach may consider such linguistic practices as *prima facie* evidence that AI robots should be taken as psychological and/or moral beings.

There are three types nonliteral interpretations. The technical view takes it that speakers have explicit intentions of expressing technical concepts when they use the relevant terms in the AI domain. The emotion view holds that the ordinary uses are essentially mediated by our own empathetic emotional states, hence the relevant terms express empathetic emotions rather than the literal meaning. According to the habit view, we are subconsciously following ingrained social habits in the ordinary uses, thus the relevant terms express such social habits, but not the literal meaning. These three views cover every previous tentatively plausible nonliteral interpretation I have come across. In course of my argument, I engage with empirical research on human-robot relations, in particular the works by Marchesi et al. (2019), Perez-Osorio et al. (2019), Ward et al. (2014), Wang and Krumhuber (2018), Rosental-von der Pütten et al. (2013), and Shin (2021). I will show that all three views are implausible with respect to the ordinary uses of agency-terms in the AI domain for they fail to account for the results of these empirical research. This constitutes my negative argument for a literal interpretation: the best
interpretation is that the ordinary uses of agency-terms constitute evidence that we (i.e., laypersons) are currently expanding the extensions of the terms to include AI robots because

there is no plausible nonliteral interpretation. On the other hand, I argue that we seem not literally considering AI robots as emotional beings or moral patients because the emotion view well explains that when we extend emotion-terms and moral-patiency-terms to AI robots, we are in fact expressing empathetic emotions.

In what follows, I first discuss the ascription of psychological states. After introducing a series of case studies about the ordinary uses of psychological terms in the AI domain, I explain each view in turn, and examine whether the interpretation accommodates experimental results. In section three, I discuss the ascription of moral patiency. At the end of each section, I discuss the extent to which the semantics of the relevant terms is preserved, or how the meaning changes, when the unexpected referential shift takes place.

References

Bennett, M., and Hacker, P. M. S. (2022). *Philosophical Foundations of Neuroscience* (second edition). New Jersey: John Wiley & Sons.

Coeckelbergh, M. (2011). You, robot: on the linguistic construction of artificial others. *AI & Soc*, 26:61-69.

Figdor, C. (2018). *Pieces of Mind: The Proper Domain of Psychological Predicates*. Oxford: Oxford University Press.

Huebner, B. (2010). Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? *Phenomenology and Cognitive sciences*, 9:133-155. Marchesi, S., Ghiglino, D., Ciardo, F., et al. (2019). Do We Adopt the Intentional Stance Toward Humanoid Robots? *Frontiers in Psychology* 10(450): 1-13.

Perez-Osorio, J., Marchesi. S., Ghiglino, D., Ince, M., and Wykowska, A. (2019). More Than You Expect: Priors Influence on the Adoption of Intentional Stance Towards Humanoid Robots. Salichs et al. eds., *ICSR 2019;LNAI*, 11876:119-129.

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., et al. (2013). An Experimental Study on Emotional Reactions Towards a Robot. *Int. J. Soc. Robot*, 5: 17-34. Shin, H. (2021). Who Has a Mind?: Mind Perception and Moral Decision Toward Robots. *Journal of Social Science*, 32:195-213.

Wang, X., and Krumhuber, E. G. (2018). Mind Perception of Robots Varies With Their Economic Versus Social Function. *Frontiers in Psychology*, 9(1230): 1-10. Ward, A. F., Olsen, A. S., and Wegner, D. M. (2013). The Harm-Made Mind: Observing Victimization Augments Attribution of Minds to Vegetative Patients, Robots, and the Dead. *Psychological Science*, 24(8):1437-1445.