No Brainer: Why Consciousness is Neither a Necessary nor Sufficient Condition for AI Ethics

David J. Gunkel[1]

[1] Northern Illinois University, USA – dgunkel@niu.edu

Abstract. The question concerning the moral and/or legal status of others is typically decided on the basis of pre-existing ontological properties, e.g. whether the entity in question possesses consciousness or sentience or has the capacity to experience suffering. In what follows, I contest this standard operating procedure by identifying three philosophical problems with the properties approach (i.e. substantive, terminological, and epistemological complications), and I propose an alternative method for defining and deciding moral status that is more empirical and less speculative in its formulation. This alternative shifts the emphasis from internal, ontological properties to extrinsic social relationships, and can, therefore, be called a "relational turn" in AI ethics.

Keywords: Artificial Intelligence, Consciousness, Ethics

1 Introduction

Ethics, in both theory and practice, is an exclusive undertaking. In confronting and dealing with others, we inevitably make a decision between "who" is morally significant and "what" remains a mere thing. These decisions (which are quite literally a cut or "*de-caedere*" in the fabric of being) are often accomplished and justified on the basis of intrinsic, ontological properties. "The standard approach to the justification of moral status is," Mark Coeckelbergh explains, "to refer to one or more (intrinsic) properties of the entity in question, such as consciousness or the ability to suffer. If the entity has this property, this then warrants giving the entity a certain moral status" [1]. According to this way of thinking—what one might call the standard operating procedure of moral consideration—the question concerning the status of others would need to be decided by first identifying which property or properties would be necessary and sufficient to have moral standing and then figuring out whether a particular entity (or class of entities) possesses this property or not. Deciding things in this fashion, although entirely reasonable and expedient, has at least three philosophical problems, all of which become increasingly evident and problematic in the face of artificial intelligence and robots.

2 Three Philosophical Problems

2.1 Substantive

First, how does one ascertain which exact property or properties are necessary and sufficient for moral status? In other words, which one, or ones, count? The history of moral philosophy can, in fact, be read as something of an on-going debate and struggle over this matter with different properties vying for attention at different times. And in this process, many properties that at one time seemed both necessary and sufficient have turned out to be spurious, prejudicial or both. Take for example the faculty of reason. When Immanuel Kant defined morality as involving the rational determination of the will, non-human animals, which did not possess reason, were categorically excluded from moral consideration. It is because the human being possesses reason, that he (and the human being, in this particular circumstance, was still principally understood to be male) is raised above the instinctual behavior of the brutes and able to act according to the principles of pure practical reason [2].

The property of reason, however, has been subsequently contested by efforts in animal rights philosophy, which begins, according to Peter Singer's analysis, with a critical intervention issued by Jeremy Bentham: "The question is not, 'Can they reason?' nor, 'Can they talk?' but 'Can they suffer?'" [3]. According to Singer, the morally relevant property is not speech or reason, which he believes would set the bar for moral inclusion too high, but sentience and the capability to suffer. In *Animal Liberation* (1975) and subsequent writings, Singer argues that any sentient entity, and thus any being that can suffer, has an interest in not suffering and therefore deserves to have that interest taken into account [4]. This is, however, not the final word on the matter. One of the criticisms of animal rights philosophy, is that this development, for all its promise to intervene in the anthropocentric tradition, still remains an exclusive and exclusionary practice. Environmental ethics, for instance, has been critical of animal rights philosophy for organizing its moral innovations on a property (i.e. suffering) that includes some sentient creatures in the community of moral subjects while simultaneously justifying the exclusion of other kinds of "lower animals," plants, and the other entities that comprise the natural environment.

But even these efforts to open up and to expand the community of legitimate moral subjects has also (and not surprisingly) been criticized for instituting additional exclusions. "Even bioethics and environmental ethics," Luciano Floridi argues, "fail to achieve a level of complete universality and impartiality, because they are still biased against what is inanimate, lifeless, intangible, abstract, engineered, artificial, synthetic, hybrid, or merely possible. Even land ethics is biased against technology and artefacts, for example. From their perspective, only what is intuitively alive deserves to be considered as a proper centre of moral claims, no matter how minimal, so a whole universe escapes their attention" [5]. Consequently, no matter what property (or properties) comes to be identified as morally significant, the choice of property remains contentious, debatable, and seemingly irresolvable. The problem, therefore, is not necessarily deciding which property or properties come to be selected as morally

significant. The problem is in this approach itself, which makes moral consideration dependent upon a prior determination of properties.

2.2 Terminological

Second, irrespective of which property (or set of properties) is selected, they each have terminological troubles insofar as things like rationality, consciousness, suffering, etc. mean different things to different people and seem to resist univocal definition. *Consciousness*, for example, is one property that has been cited as a necessary and sufficient condition for moral subjectivity [6]. But consciousness is persistently difficult to define or characterize. The problem, as Max Velmans points out, is that this term unfortunately "means many different things to many different people, and no universally agreed core meaning exists" [7]. In fact, if there is any general agreement among philosophers, psychologists, cognitive scientists, neurobiologists, ethologists, AI researchers, and robotics engineers regarding consciousness, it is that there is little or no agreement when it comes to defining and characterizing the concept. Although consciousness, as Anne Foerst remarks, is the secular and supposedly more "scientific" replacement for the occultish "soul" [8], it appears to be just as much an occult property or what Daniel Dennett calls an impenetrable "black box" [9].

Other properties do not do much better. Suffering and the experience of pain—which is the property usually deployed in non-standard patient-oriented approaches like animal rights philosophy—is just as problematic, as Dennett cleverly demonstrates in the essay, "Why You Cannot Make a Computer that Feels Pain." In this provocatively titled essay, Dennett imagines trying to disprove the standard argument for human (and animal) exceptionalism "by actually writing a pain program, or designing a pain-feeling robot" [9]. At the end of what turns out to be a rather protracted and detailed consideration of the problem—complete with detailed block diagrams and programming flowcharts—Dennett concludes that we cannot, in fact, make a computer that feels pain. But the reason for drawing this conclusion does not derive from what one might expect. According to Dennett, the reason you cannot make a computer that feels pain is not the result of some technological limitation with the mechanism or its programming. It is a product of the fact that we remain unable to decide what pain is in the first place. What Dennett demonstrates, therefore, is not that some workable concept of pain cannot come to be instantiated in the mechanism of a computer or a robot, either now or in the foreseeable future, but that the very concept of pain that would be instantiated is already arbitrary, inconclusive, and indeterminate [9].

2.3 Epistemological

As if responding to Dennett's challenge, engineers have, in fact, not only constructed mechanisms that synthesize believable emotional responses [10] [11] [12], but also systems capable of evincing something that appears to be what we generally recognize as "pain." The interesting issue in these cases is determining whether this is in fact "real pain" or just a simulation. In other words, once the morally significant property or properties have been identified and defined, how can one be entirely cer-

tain that a particular entity possesses it, and actually possesses it instead of merely simulating it? Answering this question is difficult, especially because most of the properties that are considered morally relevant tend to be internal mental or subjective states that are not immediately accessible or directly observable. As Paul Churchland famously asked: “How does one determine whether something other than oneself—an alien creature, a sophisticated robot, a socially active computer, or even another human—is really a thinking, feeling, conscious being; rather than, for example, an unconscious automaton whose behavior arises from something other than genuine mental states?” [13]. This is, of course, what philosophers commonly call “the problem of other minds.” Though this problem is not necessarily intractable, as I think Steve Torrance has persuasively argued [14], the fact of the matter is we cannot, as Donna Haraway describes it, "climb into the heads of others to get the full story from the inside" [15].

3 Thinking Otherwise

In response to these problems, philosophers—especially in the continental tradition—have advanced alternative approaches to deciding the question of moral status that can be called, for lack of a better description, “thinking otherwise” [16]. This phrase signifies different ways to formulate the question concerning moral standing that is open to and able to accommodate others—and other forms of otherness.

3.1 Relatively Relational

According to this alternative way of thinking, moral status is decided and conferred not on the basis of subjective or internal properties decided in advance but according to objectively observable, extrinsic relationships. As we encounter and interact with other entities—whether they be another human person, an animal, the natural environment, or a domestic robot—this other is first and foremost experienced in relationship to us. The question of moral status, therefore, does not depend on and derive from what the other is in its essence but on how she/he/it (and the choice of pronoun here is part of the problem) stands in relationship to us and how we decide, in the face of the other, to respond. Consequently, and contrary to the standard operating procedures, what the entity is does not determine the degree of moral value it enjoys. Instead the exposure to the face of the Other, what Levinas calls “ethics,” precedes and takes precedence over all these ontological machinations and determinations [17].

This shift in perspective—a shift that inverts the standard procedure by putting ethics before ontology—is not just a theoretical proposal; it has, in fact, been experimentally confirmed in a number of practical investigations with computers, AI, and robots. The computer as social actor (CASA) studies undertaken by Byron Reeves and Clifford Nass, for example, demonstrated that human users will accord computers social standing similar to that of another human person and that this occurs as a product of the extrinsic social interaction, irrespective of the actual intrinsic properties (actually known or not) of the entities in question [19]. These results have been verified in two

studies with robots, where researchers found that human subjects respond emotionally to robots and express empathic concern for the machines irrespective of knowledge concerning the properties or inner workings of the device [19] [20]. Although Levinas himself would probably not recognize it as such, what these studies demonstrate is precisely what he had advanced: the ethical response to the other precedes and even trumps decisions concerning ontological properties.

3.2 Radically Empirical

In this situation, the problems of other minds—the difficulty of knowing with any certitude whether the other who confronts me has a conscious mind or is capable of experiencing pain—is not some fundamental epistemological limitation that must be addressed and resolved prior to moral decision making. Levinasian philosophy, instead of being tripped up or derailed by this epistemological problem, immediately affirms and acknowledges it as the condition for possibility of ethics as such. Or as Richard Cohen succinctly describes it, "not 'other minds,' mind you, but the 'face' of the other, and the faces of all others" [21]. In this way, then, Levinas provides for a seemingly more attentive and empirically grounded approach to the problem of other minds insofar as he explicitly acknowledges and endeavors to respond to and take responsibility for the original and irreducible difference of others instead of getting involved with and playing all kinds of speculative (and unfortunately wrongheaded) head games.

This means that the order of precedence in moral decision making can and perhaps should be reversed. Internal properties do not come first and then moral respect follows from this ontological grounding. Instead the morally significant properties—those ontological criteria that we assume anchor moral respect—are what Slavoj Žižek terms "retroactively (presup)posited" [22] as the result of and as justification for decisions made in the face of social interactions with others. In other words, we project the morally relevant properties onto or into those others who we have already decided to treat as being socially significant—those Others who are deemed to possess face, in Levinasian terminology.

3.3 Literally Altruistic

Finally, because ethics transpires in the relationship with others or in the face of the other, decisions about moral standing can no longer be about the granting of rights to others. Instead, the other, first and foremost, questions my rights and challenges my solitude. This interrupts and even reverses the power relationship enjoyed by previous forms of ethics. Here it is not a privileged group of insiders who then decide to extend rights to others, which is the standard model of all forms of moral inclusion or what Singer calls a "liberation movement" [4]. Instead the other challenges and questions the rights and freedoms that I assume I already possess. The principal gesture, therefore, is not the conferring rights on others as a kind of benevolent gesture or even an act of compassion for others but deciding how to respond to the Other, who always

and already places my rights and assumed privilege in question. Such an ethics is altruistic in the strict sense of the word. It is "of or to others." This means, however, that we would be obligated to seriously consider all kinds of others as Other, including other human persons, animals, the natural environment, artifacts, technologies, and artificial intelligence. An "altruism" that limits in advance who can be Other is not, strictly speaking, altruistic.

References

1. Cocekelbergh, M. *Growing Moral Relations: Critique of Moral Status Ascription*. New York: Palgrave Macmillan (2013).

2. Kant, I. *Critique of Practical Reason*. Trans. by L. W. Beck. New York: Macmillan (1985).

3. Bentham, J. *An Introduction to the Principles of Morals and Legislation*. Oxford: Oxford University Press (2005).

4. Singer, P. *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York: New York Review of Books (1975).

5. Floridi, L. *The Ethics of Information*. Oxford: Oxford University Press (2013).

6. Himma, K. E. Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to be a Moral Agent? *Ethics and Information Technology* 11(1), 19–29 (2009).

7. Velmans, M. *Understanding Consciousness*. London, UK: Routledge (2000).

8. Benford, G. & E. Malartre. *Beyond Human: Living with Robots and Cyborgs*. New York: Tom Doherty (2007).

9. Dennett, D, C. *Brainstorms*. Cambridge, MA: MIT Press (1998).

10. Bates, J. The role of emotion in believable agents. *Communications of the ACM* 37, 122–125 (1994).

11. Blumberg, B., P. Todd, & M. Maes. No Bad Dogs: Ethological Lessons for Learning. *Proceedings of the 4th International Conference on Simulation of Adaptive Behavior* (SAB96), 295–304. Cambridge, MA: MIT Press (1996).

12. Breazeal, C. & R. Brooks. Robot Emotion: A Functional Perspective. *Who Needs Emotions: The Brain Meets the Robot*, edited by J. M. Fellous and M. Arbib, 271–310. Oxford: Oxford University Press (2004).

13. Churchland, P. M. *Matter and Consciousness*. Cambridge, MIT Press (1999).

14. Torrance, S. Artificial Consciousness and Artificial Ethics: Between Realism and Social Relationism. *Philosophy & Technology* 27(1), 9-29 (2013).

15. Haraway, D. J. *When Species Meet*. Minneapolis, MN: University of Minnesota Press (2008).

16. Gunkel, D. J. *The Machine Question*. Cambridge, MA: MIT Press (2012).

17. Levinas, E. *Totality and Infinity*. Trans. by A. Lingis. Pittsburgh, PA: Duquesne University Press (1969).

18. Reeves, B. & C. Nass. *The Media Equation*. Cambridge: Cambridge University Press (1996).

19. Rosenthal-von der Pütten, A. M., N. C. Krämer, L. Hoffmann, S. Sobieraj & S. C. Eimler. An Experimental Study on Emotional Reactions Towards a Robot. *International Journal of Social Robotics* 5, 17-34 (2013).

20. Suzuki, Y., L. Galli, A. Ikeda, S. Itakura & M. Kitazaki. Measuring Empathy for Human and Robot Hand Pain Using Electroencephalography. *Scientific Reports* 5, 15924 (2015).

21. Cohen, R. A. *Ethics, Exegesis, and Philosophy: Interpretation After Levinas*. Cambridge: Cambridge University Press (2001).

22. Žižek, S. *For They Know Not What They Do: Enjoyment as a Political Factor*. London: Verso (2008).