Journal of Ethics and Emerging Technologies

IEET

Book Review

A Review of - *Humans and Robots: Ethics, Agency, and Anthromorphism* by Sven Nyholm

Nyholm, S. 2020. *Humans and Robots: Ethics, Agency, and Anthromorphism*. Rowman and Littlefield International. 225pp. ISBN: HB 978-1-78661-226-7

Diego Morales[1]

[1] Philosophy & Ethics Group, Eindhoven University of Technology.
* Correspondence: d.h.morales.perez@tue.nl

Citation: Morales, Diego. 2023. A Review of - Humans and Robots: Ethics, Agency, and Anthromorphism by Sven Nyholm. *Journal of Ethics and Emerging Technologies* 33: 1. https://doi.org/10.55613/jeet.v33i1.122

Received: 20/01/2023
Accepted: 20/01/2023
Published: 30/06/2023

Publisher's Note: IEET stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Robots are increasingly becoming part of our social and daily lives. From robotic vacuum cleaners in our homes to sophisticated humanoid robots that greet us when we enter a restaurant or hotel, these technologies aid us in the achievement of goals, perform tasks for us, and, in some cases, can even satisfy needs. Yet, how should we behave towards robots, and how should robots conduct themselves around us are not clear and straightforward matters. In his book, *Humans and Robots: Ethics, Agency, and Anthromorphism*, Sven Nyholm delves into these questions, and aims to provide an answer that not only introduces the reader to the complexities of thinking about human-robot interactions, but also establishes a well-argued position within the field's discussions.

The book begins with an engaging analysis of its subject-matter, the ethics of human-robot interaction. From the outset, the author shows that this is a complex topic to grasp, since the meaning of the terms *ethics*, *human*, *robot*, and *interaction* is controversial, and apt to be satisfied in several, alternative ways. Stating that these terms are ambiguous might seem a truism. Yet, it continues to be a relevant starting point, given that the landscape of the field is regularly being reshaped, both in sense and scope, based on the theoretical choices we make and the notions that we employ.

For Nyholm, the aforementioned terms should be understood in the following way: *Ethics* is a normative endeavour, which studies how individuals ought to conduct themselves (p. 4). In this context, the relevant sets of individuals are those of humans and robots. While the former are described as embodied individuals that possess distinct types of bodies, minds, biological, and cultural features (p. 12), for the latter no general description is given. Instead, the author chooses to focus on particular kinds of robots that actually exist or that we may reasonably expect to build (p. 11). Finally, the interaction between humans and robots is presented throughout the book as a two-way relation between agents. As such, both relata are considered to be capable of performing actions directed at each other, which are worthy of moral consideration and regulation. Put together, then, the resulting meaning of "ethics of human-robot interaction" is the following: it is the study of the normative considerations that should govern the conduct

of humans, as embodied beings and members of a particular animal species, towards different types of actual or foreseeable robots, and the normative considerations that regulate the behavior of these types of robots towards humans.

This way of carving the ethics of human-robot interactions determines the scope of the book's proposal in two significant ways. First, regarding the subjects of the interaction. The subject-matter, as described above, assumes that there is a distinction between humans and robots robust enough to position them as opposite relata. However, this distinction and its grounds are not clearly layed out for the reader. The clues given to the reader are mainly two: (i) the suggestion that humans and robots are entities of different natures (a fact that grounds the difference in their agential capabilities – more on this later); and, (ii) that humans are characterized by a conjuction of embodiment and culture, and that actual or foreseeable robots do not satisfy this requirement. Yet, these hints are insuficient for the reader to determine whether (i) is derived from (ii), and whether (ii) is a purely contingent matter, valid only for the state of affairs at the moment of the publication of the book, or there is some *a priori* reason that impedes, and will continue impeding, robots to satisfy this characterization of human beings. Humans are unequivocally distinguishable from some *specific* kinds of robots that the author uses as examples to illustrate his claims, such as self-driving cars and autonomous weapons. For these cases, it seems unproblematic to provide such a general way of differentiating both relata. But with other kinds, such as humanoid robots of increasing embodied resemblance to us and cultural integration with us, the lines start to become blurry, making the clues provided above an insufficient demarcation criterion. Consequently, it is hard to determine whether the normative considerations put forward in the book are generalizable or they are only applicable to those cases in which the distinction between relata is clear and straightforward.

Second, concerning the forms that the interaction may adopt, as understanding the ethics of human-robot interactions as presented above widens the scope of relevant interactions beyond mere patiency. While the possibility of robots solely undergoing actions is not precluded, attributing agency increases the number and complexity of human-robot interactions that should be examined. In other words, acknowledging that robots are capable, at the very least, of initiating actions, pursuing goals, and possessing functional autonomy (p. 54) makes more complex kinds of interactions, such as collaborative agency (Ch. 3), coordination (Ch. 4), and human-robot relationships (Ch. 5), plausible and morally relevant. Moreover, as both relata possess agential capacities, the behavior of both parties is apt to be regulated by normative considerations and, therefore, susceptible to be modified in virtue of normative reasons. Therefore, it might be unreasonable to conclude that robots are always the entities that must modify, alter, or adapt their behavior whenever they engage with humans. In some forms of interaction, the author suggests, it might be morally appropriate for us to adjust our behavior towards them (p. 19).

Regarding robot agency, the core notion of the book, the author offers interesting and thought-provoking claims. Although the book does not present them in this manner, I will divide these into ontological and epistemic claims. The former concern the nature of robot agency, and the latter the way in which humans form beliefs and knowledge about said agency. On the ontological claims, readers of the book will readily infer that Nyholm endorses the view that agential capacities are determined by the nature of the agent (for example, pp. 32, 39, 201). This means that the range of actions that an entity can perform, their degree of complexity or sophistication, and the elements that accompany or precede the action (such as goal-formation, decision-making processes, and the kind of autonomy possessed) are grounded on the features that make the entity what it is. Thus, assuming that there is a clear distinction humans and robots, allows the author to infer that there is clear distinction between "robotic agency" and "human agency" as well. In light of this, the ontology of robot agency should not be understood as a mere projection of its human counterpart. It is a *kind* of agency in its own right, which should, in principle, be held to standards different from those that pertain human agency (p. 39).

Quite a different story is how humans *interpret* robotic agency. On this epistemic issue, Nyholm offers two important points. In the first place, he highlights a challenge: both human psychology and social institutions have evolved to their present state before the creation of robots (pp. 14, 35, 137). The consequence of this adaptation to robot-less environments is that humans are, presumably, poorly prepared to deal with these technologies. At an individual level, psychological features, such mind-reading (p. 16), dual processing (p. 17), tribalism (p. 18), and laziness (p. 18) ground tendencies to anthropomorphize robots. At an institutional level, our legal systems include operative notions of agency that do not perfectly overlap with that of "robotic agency". For example, these definitions usually require that agents act freely and spontaneously in order for their actions to have legal validity; elements which, in principle, robots don't seem able to exhibit. In virtue of these features, humans are prone to either distort and misinterpret robotic actions by attributing to them specific goals, desires, and intent, or to assumine that they play no agential or active role in legal discussions about responsibility.

In the second place, the author offers a methodological suggestion: robot agency, being of a different kind from human agency, may be challenging for us to grasp, and our need to interpret it may be inescapable. However, "we should seek ways of interpreting and talking about *robots' actual capacities* in ways that permit the language of agency" (p. 42, my emphasis). This means that we should not relinquish our interpretative tendencies, but we must exercise them with some caveats in mind. We should, for example, engage in talk about the actual agential capabilities of concrete types of robots rather than in talk oriented towards looking for necessary and sufficient conditions for all robot agency. And this is precisely the choice followed in the book, as agency and its implications are discussed through particular cases, such as self-driving cars, autonomous weapons, and sex robots.

All this groundwork, while providing a plausible framework to understand the ethics of human-robots interactions, faces challenges when it comes down to the details. One puzzling aspect lies in the relation between the general notion of agency and particular notions of the term (e.g. human agency and robot agency). The book says nothing about whether the general notion of agency and particular variants are related as genus and species, or as determinable and determinates. This rather theoretical discussion is not inconsequential for the book's proposal. As human beings must interpret robot agency, the way in which they ought to do it will be impacted by how the relation between concepts is understood. On the one hand, if we approach the issue with a genus-species model in mind, interpreting robot agency becomes the quest to determine which features are shared with human agency (in virtue of being derived from the same *genus*, namely, "general agency") and finding the unique *differentia* that picks out this concept from the other *species* concepts (e.g. human agency or non-human animal agency). The same exercise must be iterated between the notions of robot agency and its subspecies, such as self-driving cars agency or sex robot agency. On the other hand, if we endorse a determinable-determinate model, then interpreting robotic agency amounts to understanding that it is a determinate of the general notion of agency, but that no separable *differentia* uniquely picks it out. Rather, it is a kind of agency that is similar, yet incompatible with other determinates, such as human agency. This means that ROBOT AGENCY is a concept that cannot be realized alongside other concepts of agency by the same entity at the same time, but the degree of similarity that shares with them grounds the resemblance relation that causes so many interpretative headaches. Unfortunately, the book moves ambiguously between these theoretical options, at the expense of providing more clear and solid grounds for its proposal.

Another aspect in which the book could have benefitted from more precision concerns agency *attribution*. As mentioned above, robot agency is to be considered a kind of agency in its own right, to be held to its own standards, and justified on independent grounds from those of human agency. As such, the book's framework suggests that our first interpretative task should be to acknowledge that robot agency has these features. If no normative implication were to be derived from this acknowledgement, then the notion of *attribution*, or mere ascription of possession, would be adequate to denote our interpretative activities. However, Nyholm points out that, at least in the case of humanoid robots, certain behavior is expected from us: we ought to treat them in respectful, dignified, or considerate ways due to the agency that they exhibit (p. 187). If our acknowledgement of agency is to be accompanied by normative considerations, then the richer notion of *recognition* would do a better explanatory job that the one of *attribution*. This is because *recognition* denotes a complex act by which a certain capacity or property is regarded as possessed, instantiated, or realized by a given entity, and an obligation is acquired to treat said entity in a certain way, in virtue of said capacity. Typically, then, *recognition* yields the consideration that an entity possesses a specific normative status.

Regardless of these comments, Sven Nyholm achieves an insightful book that accomplishes the double task of providing a general framework and an application to case

studies to study the ethics of human-robot interactions. The present work, I believe, might be of great use and interest of a wide range of audiences. Beginners in the field will benefit from a detailed scholarly work in the form of a considerable collection and commentary of arguments on several aspects of the subject-matter. More experienced researches will find a set of rigorous, engaging, and well-argued positions on open and controversial discussions. Professionals in other fields, such as law and legal practice, may find in this book an informative and accessible introduction to foundational aspects of problems that also concern their fields. And, finally, it may constitute a pleasant and instructive read for anyone with an interest in how to interact with the robots amongst us.

Acknowledgements

Special thanks to María Jesús Parga, Fabio Tollon, Céline Budding, and Patrik Hummel for their valuable feedback and comments on previous versions of this review. Work on this review was supported by funding from the TU/e – EAISI Doctoral Position for Project "AI Planner for the Future".