Sven Nyholm: Humans and Robots: Ethics, Agency, and Anthropomorphism

Sand, M.

DOI
10.1007/s10677-020-10083-2

Publication date
2020

Document Version
Accepted author manuscript

Published in
Ethical Theory and Moral Practice

This manuscript is pre-edited version of a paper that has been published in Ethical Theory and Moral Practice. Please refer to the published version: doi: 10.1007/s10677-020-10083-2.

Sven Nyholm: Humans and Robots: Ethics, Agency, and Anthropomorphism

***

Review: Nyholm, Sven (2020): Humans and Robots: Ethics, Agency, and Anthropomorphism. Rowman & Littlefield International. Series: Philosophy, Technology and Society. ISBN: 9781786612274. 169 Pages. 23,95 £.

***

With "Frankenstein, or the modern Prometheus", Mary Shelley has masterfully expounded the question of artificial agency and responsibility for autonomous technologies that regularly re-emerges in philosophical and public debates ever since (Johnston, 2016). Because of recent advancements and the increasing diffusion of semi-autonomous cars, military technologies and intelligent algorithms and the accidents that result from them, the topic becomes more exigent than ever. In "Humans and Robots: Ethics, Agency, and Anthropomorphism", Sven Nyholm presents a comprehensive treatment of the subject focusing on two ethical questions: "On the one hand, how should robots be made to behave around people? On the other hand, how should people conduct themselves around different kinds of robots?" (p. **4**) The book, however, frequently exceeds the perimeter of these questions, for instance, when Nyholm asks whether robots can have "humanlike" minds and whether they can be meaningfully conceived of as moral agents, which he both denies (p. 146, 160).

In the first chapters, Nyholm introduces the concepts and premises essential to his subsequent reflections. An important presupposition introduced early on, suggests that we must not only ask which kinds of robots are desirable in the future but also which kind of behavior might be demanded from humans given the

compelling reasons to utilize autonomous technologies in various contexts (p. 20). Taking this premise into account, Nyholm defends to mandate speed-regulation in ordinary cars to reduce the risk for traffic incidents in mixed-traffic situations in chapter 4 (p. 89). In the first two chapters, Nyholm also endorses more substantive anthropological theses. He suggests that human minds, which evolved before the advent of autonomous machines, are somehow "unfit" to interact with robots: "So, just like I argued in chapter one that our minds are not necessarily well-adapted to interact with robots and AI, I also wish to suggest that our legal and ethical doctrines – and, along with them, our ideas about agency – are not necessarily well-adapted to deal with robots and AI." (p. 35) While Nyholm emphasizes an affinity to Persson's and Savulescu's thesis of humanities "unfitness for the future" (Persson & Savulescu, 2012), the reader might also detect similarities to the presuppositions of Hans Jonas' "Imperative of Responsibility". In the wake of nuclear weapons and the lasting effects of modern technologies on the environment, Hans Jonas suggested that previous ethical theories are insufficient in an age, where technology has altered the "nature of human action" (Jonas, 1984, p. 1). The comparison to Jonas lends itself most naturally when Nyholm states that "human-robot interaction raises philosophical questions that require us to think creatively and innovate ethical theory." (p. 6) However, unlike Jonas, who developed a new ethical theory to fill an alleged ethical vacuum (p. 23), Nyholm pursues a more modest path, often drawing on existing ethical theories and relying on "widely shared ethical ideas." (Fn. 15, p. 23) Given this approach, one wonders whether his initial commitment to a number of (debatable) anthropological assumptions about humans' (in-)ability to coexist with robots were necessary. Most readers would have been convinced of the book's urgency (the conclusion drawn on

page 16) based on the obvious challenges that robots pose to our current ethical and legal frameworks (p. 35).

Nyholm doubts that the question, whether robots are agents, can always be clearly answered (p. 31). He believes that this question is normative and related to concerns about responsibility allocation. Here, the widespread tendency to anthropomorphize robots leads him to caution of rejecting robot agency right away: He adopts a stance that "instruct[s] us to try to find acceptable ways of interpreting robots as some sorts of agents." (p. 42) But, is this so-called "moderate conservatism" actually defensible and what if the only acceptable way of interpreting a robot's behavior is indeed the denial of its agency *contra* many people's inclination?

As said, Nyholm often calls existing ethical theories in to assess robot-human interaction: In chapter 5 (p. 120), he draws on Cicero's theory of friendship to justify, why robots cannot be "real" friends and in chapter 8, Kant's "formula of humanity" backs the claim that one might have to treat robots respectfully, not for their own sake but for humanity's sake (p. 187). In chapter 4, where Nyholm focusses on the ethical implications of autonomous vehicles, he utilizes the child-parent analogy to shed light on the issue of responsibility for autonomous technologies' failures. Based on this discussion, he classifies different types of human-robot collaborations depending on the various ways in which humans and robots interact. This results in a more dynamic model of collaborative agency and allows for a more concise identification of "responsibility-loci" based on various considerations including; who has supervision control, who switches the technology on, who understands its functioning and monitors its behavior. In sum,

this chapter presents a convincing case to overcome conceiving robot agency as a binary matter.

In chapter 8, which is concerned with the moral status and rights of robots, Nyholm presents an argument for treating robots respectfully "for the sake of humanity." There, the reader will likely miss a more definite viewpoint. He suggests that Kant's "formula of humanity" "could […] require us to treat the robots that perhaps already exist or that will soon exist with some degree of respect and dignity." (p. 189) *Could* the principle require or does it in fact require some degree of respect? The chapter raises a more general worry: Non-Kantians will search without much success for a reason to accept the "formula of humanity" on which much of the chapter's conclusions are based.

All criticism aside, this comprehensive book on robot ethics is written with great care and clarity. Readers without previous knowledge of the subject will find Nyholm's recapitulations of the most recent literature instructive and its untechnical style accommodating. The chapters often start out with topical anecdotes that underscore the connection between the philosophical debate and the real-life consequences of the increasing implementation of artificially intelligent technologies. Nyholm reads his opponents favorably and refrains from creating strawmen, which is additional evidence of his proficiency. What some will consider a strength, others will find this book's greatest weakness: Despite the "existential" tension between human nature and robots evoked in the first chapter, Nyholm – unlike Jonas before him – does not turn the field of ethics inside out. Instead, he defends numerous more confined propositions that deserve careful inspection.

Those who think that the robot challenge requires a more radical approach – a new "imperative of robot ethics" – might be disappointed. Others who have often been underwhelmed by pompous proposals that eventually fell short of substance and coherence, will find great merit in Nyholm's unagitated and balanced reflections.

References

Johnston, J. (2016). Traumatic responsibility: Victor Frankenstein as creator and casualty. In D. H. Guston, E. Finn, & J. S. Roberts (Eds.), *Frankenstein* (pp. 201-207). Cambridge, Mass.: MIT Press.

Jonas, H. (1984). *The Imperative of Responsibility.* Chicago & London: University of Chicago Press.

Persson, I., & Savulescu, J. (2012). *Unfit for the future: The need for moral enhancement* (1 ed.). Oxford: Oxford University Press.

This manuscript is pre-edited version of a paper that has been published in Ethical Theory and Moral Practice. Please refer to the published version: doi: 10.1007/s10677-020-10083-2.