

**Book review of Seibt, J., Hakli, R., & Nørskov, M. (Eds.). (2026).  
Robophilosophy: Philosophy of, for, and by social robotics.  
Cambridge, MA: MIT Press. 489 pp. ISBN: 9780262361699.**

Reviewed by

Bogdan-Constantin Ibanescu  <sup>1</sup> 

**Type:** Book review

**Citation:** Ibanescu, B.-C. (2026). Book review of Seibt, J., Hakli, R., & Nørskov, M. (Eds.). (2026). Robophilosophy: Philosophy of, for, and by social robotics. Cambridge, MA: MIT Press. 489 pp. ISBN: 9780262361699. *ROBONOMICS: The Journal of the Automated Economy*, 7, 96

---

<sup>1</sup> Researcher, Centre for European Studies, Faculty of Law, Alexandru Ioan Cuza University of Iasi, 700506 Iasi, Romania;  
Email: [ibanescu.bogdan@uaic.ro](mailto:ibanescu.bogdan@uaic.ro)

 Corresponding author

---

**Publication history**

Received: 07/04/2026; Revised: 23/04/2026; Accepted: 28/04/2026; Published online: 29/04/2026; Volume date: 31/12/2026



© 2026 The Author(s)

This work is licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0).

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

---

At first glance, *Robophilosophy* might appear to be an eccentric scientific undertaking. For some disciplinary purists, even its title may seem almost oxymoronic, as if robotics and philosophy belonged to separate epistemic worlds. Yet reading this volume quickly dispels that impression. There is nothing eccentric here, only the necessity of confronting a reality that is already unfolding around us. Whether or not we fully acknowledge it, we are already living amid the questions that this book addresses. Social robots are no longer confined to speculative futures or laboratory curiosities; they are part of a broader technological transformation that is beginning to alter social relations, moral frameworks, and understandings of human agency.

It should be acknowledged from the outset that, despite every effort to keep this review within a conventional length, the complexity of the book and the scope of its implications demand a more extended engagement. This is not simply because the volume is rich in content, but because it addresses issues that are difficult to contain within a detached summary. The reflections it provokes are not external additions to the reading experience; they emerge from within it. For this reason, some personal reflections are interwoven throughout the present review, not as digressions, but as a response to the book's own intellectual ambition.

The volume brings together fifteen excellent chapters that explore both the internal and the relational dimensions of philosophy as they pertain to social robots. By the end of the book, the reader is unlikely to leave with a fixed, ready-made position. This is one of its strengths. The volume does not seek to function as a plea for one scenario or another, nor does it advocate a single ideological stance on robotics. Rather, it offers a more rigorous and nuanced understanding of what may be expected in the coming decades, while also underlining the significance of the decisions that are already being taken in the field of robotics and artificial intelligence.

From the very beginning, the book establishes its terms with remarkable clarity. It does not merely define robophilosophy; it also states its purpose explicitly, namely, to offer “a systematic introduction to robophilosophical research”. This promise is delivered admirably. The volume is not written for speed readers. Nearly every paragraph is carefully weighed, and almost every page opens conceptual connections to existing debates, theoretical traditions, empirical problems, or philosophical tensions. It is a demanding book, but it is demanding in a productive way.

Although the volume consists of fifteen chapters grouped into three parts, it never reads as a loose collection. The three sections are closely related, and at first sight one might even be tempted to think they differ only slightly, especially since their titles vary by only one word. Yet these distinctions matter, and the book rewards slow reading precisely because it insists on conceptual precision. The chapters are consistently strong, carefully documented, and often so elaborated that some of them could almost stand alone as short books.

This is also one of those volumes in which the introduction should not be skipped. Indeed, *What Is Robophilosophy?*, by Johanna Seibt and Raul Hakli may be one of the most important parts of the book. As an introduction ought to do, it provides the gateway to everything that follows, but it does so with a degree of conceptual richness that many introductions no longer attempt. The notion of the social robot is central here and is explored from multiple angles. The editors also make a significant methodological point: in order for robophilosophy to address its core questions, dilemmas, and exploratory tasks adequately, it must move beyond some of the standard methodologies of philosophy. One might add that it must also move beyond some of the inherited methodological assumptions of robotics itself. The volume defines robophilosophy as the philosophy **of, for, and by** social robots, and this tripartite formulation is not a rhetorical device; it structures the whole book.

The first part, **Philosophy of Social Robotics** (Chapters 2 to 5), deals with reflective and normative responses to the phenomena generated by robots. It explores the implications of robotics for interpersonal relations,

social practices, and broader forms of cultural self-understanding. While it includes concerns usually associated with robo-ethics, it extends beyond them into ontology, metaphysics, and political philosophy.

This first section is particularly successful in showing that social robotics is not merely a technical domain to which ethics can be externally applied. Rather, it is a field that forces ethical thought itself into reconsideration. Chapter 2, by David Gunkel (*The Machine Question: Rethinking Moral Philosophy in the Face of Others*) begins from the provocative claim that we, too, are robots in certain respects, before turning to the central issue of moral standing: by what criteria do we determine which beings deserve moral consideration? The discussion navigates questions of properties, reason, pain, belief, and personhood, while also reminding the reader that human history itself is full of examples in which moral status was distributed unevenly and often violently. The critique of the traditional “properties approach” to moral standing is especially compelling. It shows how unstable our own ethical frameworks become when confronted with non-human entities designed for social interaction.

Another important contribution of this section is the reversal of perspective regarding the social. Instead of asking only what we should think about robots, Chapter 3 by Mark Coeckelbergh asks what robots can teach us about the social itself (*The Automation of the Social: What Robots Can Teach Us About the Social*). This shift is subtle but profound. Robots are no longer treated simply as objects of moral concern, but as mirrors and mediators through which human beings come to rethink their own categories of relation, communication, and sociality. This is one of the most illuminating moves in the book, because it resists reductive understandings of the social as something fully captured by operational parameters or measurable criteria.

The section also includes a stimulating discussion of nudging and robots in Chapter 4 by Raffaele Rodogno (*Who Is Afraid of Nudging by Robots?*). Nudging is well established in the social sciences, especially in behavioral economics and social psychology, but its relationship to robotics has received less sustained philosophical attention. Here, the ethics of robotic nudging are examined with nuance. The argument is not that all forms of robotic influence are automatically violations of autonomy, but that their legitimacy depends on their relation to the user’s own ends and on the broader problem of agency. Particularly striking is the suggestion that technologies without moral agency may nonetheless become highly effective agents of behavioral influence. The result is a chapter that leaves the reader productively unsettled, aware at once of opportunities and risks.

The final chapter of the first section, Chapter 5 by Illah Reza Nourbakhsh, is dense and wide-ranging (*Robots, Empowerment, and Equity*). It touches on technological, social, and moral dimensions while bringing the discussion closer to debates that are already visible in broader public discourse: human-robot interaction, hierarchy, responsibility, and especially decision-making in contexts marked by complexity and high stakes. One of the chapter’s central arguments is that decisions in robotics should not be left to non-experts who are unable to grasp the full implications of these systems, particularly the difference between short-term utility and long-term societal impact. Equally important is the call for communication by experts capable of explaining both the promises and the dangers of technological development, beyond alarmist simplifications or uncritical enthusiasm.

The second part, **Philosophy for Social Robots** (Chapters 6 to 10), shifts the emphasis from reflective assessment to conceptual construction. Here philosophy acts as a constructive partner for engineering and human-robot interaction research. It offers conceptual tools, descriptive frameworks, and heuristics that can support both robotic design and empirical inquiry.

This is perhaps the most methodologically oriented part of the book, and it is especially valuable for readers interested in how philosophical work can intervene upstream in technological development rather than merely commenting on it afterward. In Chapter 6 by Raul Hakli (*Taking a Social Stance Toward Robots*) argues

persuasively against extending our existing concepts too far when describing robots and their interactions with humans. At the same time, it questions whether our inherited concepts of the social and the moral are themselves adequate to contemporary developments. The implication is clear: the field may need not merely revised definitions but new terminology altogether.

A particularly strong contribution in this section is the development of OASIS in Chapter 7 by Johanna Seibt, Christina Vestergaard, and Malene Flensburg Damholdt (*OASIS: A Human-Centered Descriptive Framework for Human–Robot Interactions*). This framework is designed to address the “description problem” by offering a non-metaphorical vocabulary for describing robotic capacities and human-robot interaction. The notion of “sociomorphing” is especially productive here, as it captures the process through which humans attribute social capacities to robots without collapsing those capacities into simple anthropomorphism. What makes this chapter especially noteworthy is its insistence that first-person experience in interaction is not epistemically trivial. The human experience of a robot is treated as ontologically significant, even if that experience does not map neatly onto the robot’s underlying technical architecture.

Other chapters in this section pursue even deeper ontological questions. Chapter 8 by Mark Bickhard asks whether artificial agents can genuinely care or participate in real social relationships (*Robot Sociality?*). Rather than treating sociality as a matter of complex routines, the argument focuses on care, historicity, and stake. Relationships are presented not as fixed states but as temporal processes. This chapter is among the most philosophically demanding in the volume and may be less immediately accessible to non-specialists, but it is also one of the most rewarding because it reveals how much of what appears self-evident in public debates about AI and robots is in fact conceptually underdeveloped.

Chapter 9 by Jens Christian Bjerring and Jacob Busch (*Dispersed Robots and Intelligent Environments: A Philosophical Perspective*) explores the idea of dispersed robots and intelligent environments. It challenges the conventional view of the robot as a stand-alone entity and proposes instead a distributed model in which sensing and computation are embedded in a smart environment. This is one of the chapters that most forcefully pushes the reader to abandon familiar images of robotics and to think instead in systemic, relational terms. It is not only a philosophical argument; it is also an intervention into the dominant imaginaries of the field.

The final chapter of this section, Chapter 10 by John P. Sullins, addresses automated ethical practical reasoning (*Automated Ethical Practical Reasoning: The Problem of Artificial Phronesis*). At a time when advances in AI occur with remarkable speed, this chapter feels particularly urgent. Its engagement with phronesis makes for demanding reading, yet its central proposition is compelling: ethics is a form of principled experimentation, and practical wisdom is a constituent part of meaningful intelligence. If some parts of the discussion will inevitably date as the field evolves, the conceptual stakes it identifies are likely to remain.

The third part, **Philosophy by Social Robots** (Chapters 11 to 15), marks another significant shift. Here social robotics becomes a means of gaining insight into the human condition itself. Robots are no longer considered only as objects of analysis or as recipients of conceptual tools; they become instruments through which empirical and philosophical inquiry can reconsider what defines human beings, human conduct, and human sociality. I choose to review this final part more as a block than chapter by chapter, not out of lesser regard for its individual contributions, but because it reads with greater thematic continuity and perhaps with slightly more narrative ease than the preceding sections.

Taken together, these chapters show how social robots can illuminate morality, common sense, simulation, interpretation, responsibility, and the boundaries of the human. Chapter 11 by Benjamin Kuipers (*Ethical Common Sense for Robots*) highlights a critical tension: robots do not follow human intentions in the rich and

ambiguous way that humans interpret one another; they follow instructions and rule sets. This is more troubling than it first appears, because human beings are not especially good at formulating exhaustive instructions, particularly for systems that are radically unlike us. Chapter 12 by Domenico Parisi and Stefano Nolfi, in a more speculative register, mirrors humans and computers through the notion of the simulated human being (*Robotics as a New Science of Human Beings*). It has, in places, the texture of science fiction, yet this should not be read as a weakness. On the contrary, it underscores how deeply robophilosophy is entangled with long-standing cultural imaginaries. Chapter 13 by Selmer Bringsjord, Alexander Bringsjord, and Naveen Sundar Govindarajulu engaging Edgar Allan Poe is one of the most surprising in the volume (*What Would Poe Say About Today's Social Robots?*). It connects the dilemmas of robotics to literary questions of interpretation, detection, and mind-reading. For readers unfamiliar with the depth of Poe's relevance here, the chapter performs a double function: it opens a path into robophilosophy while also inviting a rediscovery of a major literary figure. Elsewhere, the experimentally grounded Chapter 14 by Peter H. Kahn, Jr. (*Consciousness, Authenticity, and Transcendence of Social Robots*) shows that humans already attribute moral standing and moral responsibility to robots. This is a significant result, not least because it suggests that forms of social and even intimate attachment to robots may emerge sooner than many would expect. Such a possibility may seem exaggerated if one thinks only of advanced humanoid machines, but it appears much less far-fetched when considered in light of present-day human-AI attachment. The Chapter 15 by Marco Nørskov, Ryuji Yamazaki-Skov, and Hiroshi Ishiguro (*Android Philosophy and Android Literacy*), deepens these concerns further and acts in some respects as a continuation of the questions raised earlier about human-robot interaction, projection, and the limits of anthropomorphic imagination.

Overall, *Robophilosophy* is an exceptionally rigorous and ambitious work. One of its greatest strengths lies in its refusal to rely on facile comparisons between social robots and earlier technologies such as the engine, the computer, or the internet. The editors argue convincingly that social robots constitute a distinctive development because they are designed not merely as tools, but as social others. This marks an important discontinuity in the history of artefacts and, correspondingly, requires conceptual responses that are equal to that novelty.

The volume's commitment to philosophical pragmatism is evident throughout. At the same time, it does not avoid darker possibilities. The book is careful not to become either excessively optimistic or excessively pessimistic about the future. Instead, it acknowledges both beneficial and troubling scenarios as real possibilities and encourages preparedness rather than ideological comfort. In that sense, the volume is intellectually sober. It takes the field seriously without surrendering to either technological euphoria or apocalyptic rhetoric.

One of the few limitations of the book is one common to many collective volumes. Certain concepts recur across chapters with only slight variation, and this can occasionally create the impression of repetition. At times, such overlap reinforces the coherence of the volume and allows ideas to resonate across different approaches. At other times, however, it may slightly affect the reading flow, especially for readers less familiar with the philosophical terrain. This is already a demanding book, and some thematic overlap can make the progression feel heavier than necessary. For that reason, it is perhaps best read slowly, one chapter at a time, allowing the arguments to settle and the implications to unfold gradually.

Yet even this relative weakness is inseparable from the volume's broader strength: its density. This is not a book built around intellectual shortcuts. It expects sustained attention and rewards it generously. For researchers in social sciences as myself, this particular aspect is extremely important and plays a significant role in the overall approach. Much current discussion of AI and robotics remains trapped either in technical vocabularies that underplay social meaning or in media framings that exaggerate novelty without conceptual depth. *Robophilosophy* offers a valuable alternative. It helps place robotics within larger discussions of normativity, power, sociality, knowledge, and the human condition. In doing so, it also reminds social scientists that robotics

is not only a matter for engineers and philosophers, but a field with far-reaching implications for the analysis of institutions, interactions, inequalities, and future forms of coexistence.

In conclusion, *Robophilosophy: Philosophy of, for, and by Social Robotics* is essential reading for anyone seriously interested in the future of AI, robotics, or philosophy itself. It provides a comprehensive and intellectually demanding response to what may rightly be called the robotic revolution. More than a critique of what may be transformed or lost, it offers a way of thinking about how humans might consciously shape their co-evolution with machines. As the editors suggest, robophilosophy will remain relevant as long as it remains an open question whether the task of understanding the mind can be undertaken by robots.

A final remark may be added. At the moment of writing this review, an increasing number of analyses converge on the view that the recent boom in AI will have major repercussions for robotics. A comparable acceleration in robot-based innovation now seems likely. The signals are persuasive. If this is indeed the direction in which the field is moving, then one may only hope that those designing, funding, regulating, and celebrating that future will also take the time to read this book. It would not slow innovation down. It might simply make it wiser.