Studies in the Philosophy of Sociality 9

Raul Hakli
Johanna Seibt *Editors*

Sociality and Normativity for Robots

Philosophical Inquiries into Human-Robot Interactions

Springer

Studies in the Philosophy of Sociality

Volume 9

More information about this series at http://www.springer.com/series/10961

Raul Hakli • Johanna Seibt
Editors

Sociality and Normativity for Robots

Philosophical Inquiries into Human-Robot Interactions

Springer

Editors

Raul Hakli
Department of Political and Economic Studies
University of Helsinki
Helsinki, Finland

Johanna Seibt
Research Unit for Robophilosophy
Department for Philosophy and History of Ideas
School of Culture and Society
Aarhus University, Aarhus, Denmark

Studies in the Philosophy of Sociality
ISBN 978-3-319-53131-1 ISBN 978-3-319-53133-5 (eBook)
DOI 10.1007/978-3-319-53133-5

Library of Congress Control Number: 2017938836

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Social robots, if used pervasively in society, will change the fabric of human social interactions more profoundly than any other technology before – or so we have currently reason to believe. Since human social interactions realize human values, any change in social interactions also potentially affects the network of our sociocultural and ethical values. That social robotics presents us with far-reaching ethical questions has been observed at least a decade ago. Since this time, however, it also has become clear that the way in which humans react to social robots does not allow for normative assessments using common subsumptive methods of applied ethics. The empirical facts of human interactions with social robots show that these interactions cannot be conceived in familiar categories of human-machine interaction or human-computer interaction. Rather, it appears that we need to explore whether the notions of sociality and normativity, the hallmarks of human-human interactions, can be suitably extended to capture the phenomena of human interactions with so-called "social" robots. Only in tandem with addressing the descriptive tasks of social robotics can adequate ethical assessments be formulated.

Thus social robotics presents us with a formidable challenge in descriptive as well as in ethical regards. In answer to this challenge, a new area of philosophy has constituted itself which aims to come to terms with the very idea of artificial social agency – "robophilosophy." The articles we have collected here – some of which have grown out of earlier articles presented at the inaugural 2014 event of the Robo-Philosophy series (Seibt, J., Hakli, R., and Nørskov, M., eds.: *Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014*, IOS Press) – for the first time take up the narrowly conceptual-analytical tasks introduced by *social* robotics. While human-computer interaction has been investigated with broad interdisciplinary scope including philosophy, the interdisciplinary field of HRI (human-robot interaction) largely still proceeds without the contributions of humanities research in general and of philosophy in particular.

The main aim of this book is to demonstrate that robo-ethics is not enough – we need to engage all philosophical disciplines, in particular also theoretical disciplines such as ontology and philosophy of mind, in a joint interdisciplinary endeavour of coming to terms with the descriptive and normative tasks social robotics is putting before us. Even though the contributions to this book address the research community in philosophy, they are accessible to an interdisciplinary readership; thus the book also lends itself as textbook for graduate courses on human-robot interaction taught from a broad range of different disciplinary angles.

We thank all those who made the book possible. Most of all, we would like to thank the authors of this volume for their contributions and their patience with the production process. But we also wish to thank Marco Nørskov, co-organizer of the Robo-Philosophy 2014 conference, who has been promoting philosophy of social robotics at the Aarhus Philosophy Department since 2008; it is ultimately due to him that we, jointly with other philosophers in Aarhus, began to explore the rich systematic perspectives of this field. We would like to thank all members of the "Research Group for Philosophical Enquiries in Social Robotics" (PENSOR group, 2012–2015) and its successor, the "Research Unit for Robophilosophy," for many fruitful discussions: Raffaele Rodogno, Jens-Christian Bjerring, Jacob Busch, Ryuji Yamazaki, Stefan Larsen, Malene Flensborg Damholdt, Christina Vestergård, Glenda Hannibal, Thea Puggaard Strand, Rikke Mayland Olsen, and Søren Schack Andersen. We are grateful to Arto Laitinen for reading and commenting on the first draft, and we thank the series editor Raimo Tuomela for admitting into the series a new perspective on the current and future significance of social ontology research. Finally, we gratefully acknowledge the support of the Velux Foundation for the research project "Social Robotics and Human Self-Understanding" (2012–2015), in the context of which this book was produced.

Helsinki, Finland
Aarhus, Denmark
September 2016

Raul Hakli
Johanna Seibt

Contents

Chapter 1
"Sociality and Normativity for Robots": An Introduction

Raul Hakli and Johanna Seibt

Official projections predict that in the course of the next two decades human societies will see pervasive use of robotic technology in all contexts of social interaction, public and private. Currently, so-called "social robots" are designed for use in elderly care and education, with applications ranging from cognitive training to language tutoring to autism therapy and research. Social robots are also used as "tour guides", "diet coaches", "personal assistants", and "cleaners", and it is predicted that by 2020 they will "influence every aspect of work and home" (euRobotics aisbl, 2013).

The accuracy of these predictions remains to be seen, but it is safe to say that there will be changes that can profoundly affect our cultural practices and social relationships. Of course, technological change has taken place throughout the history of human kind, and in people's experience the rate of change is often seen as increasing and its effects as more dramatic than before. During the past three decades we have experienced in tight sequence how the introduction of new technology has changed our everyday lives. The first wave of change occurred when personal computers entered our workplaces and our homes; the second and even farther reaching change in our social practices set in when mobile phones and then smart phones delocalized communication and information, making us accessible nearly everywhere. The ubiquitous modifications of "information technology (IT)" in the first phase were compounded by the impact of what we came to label

R. Hakli (✉)
Department of Political and Economic Studies, University of Helsinki, P.O. Box 24
(Unioninkatu 40A), FI-00014, Helsinki, Finland
e-mail: raul.hakli@helsinki.fi

J. Seibt
Research Unit for Robophilosophy, Department for Philosophy and History of Ideas,
School of Culture and Society, Aarhus University, Denmark Jens Chr. Skous Vej 7,
DK-8000, Aarhus C, Denmark
e-mail: filseibt@cas.au.dk

R. Hakli, J. Seibt (eds.), *Sociality and Normativity for Robots*, Studies in the
Philosophy of Sociality 9, DOI 10.1007/978-3-319-53133-5_1

"information and communication technology (ICT)" in the second phase. Now we are on the brink of possibly even more "disruptive" reconfigurations of socio-cultural space due to what could be called the "information, communication, and robotic technology (ICRT)" on the horizon. Arguably, the proliferation of the use of robotic technology will have even greater effects than the previous two phases since robotic "social" agents put the new powers of information and communication technology into physical space, and thereby not only extend our social interaction space, but also structure and delimit it in novel ways.

These prospects call for philosophical reflections at two levels. On the one hand, there are ethical questions concerning specific robotics applications, especially if these replace human interaction partners. Until now concerns in the area of applied ethics ("roboethics") have dominated philosophical discussions involving social robotics. On the other hand, social robotics raises also fundamental conceptual and metaphilosophical issues that philosophers need to address in addition to or even before answering ethical questions about specific applications. Social robotics technology is a far cry from the mechanization of our physical environment that we know from vending machines and automated cash counters in supermarkets – in fact, as has been noted variously in the new area of "Human-Robot Interaction" (HRI) research, we do not yet possess the right categories to classify this new sort of social agent and our interactions with it.

There is something deeply unsettling about the concept of a social robot, both unnerving and enticing in its possibilities. Social robotics is not only the engineering of robotic movements, it is the engineering of human social actions. So far we have been shaping the actions of our fellow human beings by means of ideas, rules, norms, incentives, physical objects, and environments, but none of what was human made could engender a social interaction – the only sort of item that could elicit a social response was another human agent, something that belonged to the sphere of sociality but was not manufactured. The unsettling aspect of social robotics lies in the fact that for the first time humans manufacture the kind of item that affords social interactions, or so it appears.

So-called "social" robots are designed to engage us in social interactions and to enable us to develop social relationships with them, and as empirical research shows, humans are quite willing to accept them as social partners. This is in conflict with our traditional conceptions of sociality and social interaction, which presuppose the essential capacities of human subjectivity (consciousness, rationality, intentionality, free will, normative agency etc.) in all partners of a social interaction. In short, the phenomena of human interactions with "social" robots contradict the longstanding notion that sociality is grounded in reciprocity. Since reciprocity-based conceptions of sociality are at the centre of the Western understanding of human nature, of our moral standing, our political authority, and our human rights, the very idea of artificial social agents, together with its empirical validations, presents a formidable challenge to Western philosophy.

In answer to this challenge the new area of philosophy is constituting itself which aims to come to terms with the very idea of artificial social agency – "robophilosophy" (Seibt, Hakli, & Nørskov, 2014). Robophilosophy is in most cases an interdisciplinary research effort in the intersection of philosophy, robotics, computer science, cognitive science, anthropology, psychology, and sociology. Robophilosophy is philosophy of, for, and by social robotics – it examines the socio-cultural implications of inserting a new type of agents into the space of human social interaction; it develops new conceptual tools that will help us to understand human-robot interactions and program robots in accordance with these; and it is experimental philosophy, using robots as a new investigative tool to explore the conditions of human social interaction (Seibt, Hakli, & Nørskov, 2018; Seibt, 2017).

This volume is a contribution to robophilosophy that builds a particularly important "access road" into this new field. By now several explorations into the conceptual and cultural implications of social robotics from a largely "continental" perspective in philosophy (Coeckelbergh, 2012; Gunkel, 2012; Nørskov, 2015) and "roboethics" and "machine ethics" have already staked their claims (e.g. Sullins, 2005, 2006; Sharkey & Sharkey, 2012; Wallach & Allen, 2009), researchers in analytical philosophy, and especially in social ontology, have yet to find their way into philosophy of social robotics. The present volume aims to create this important link, and to show that it is to mutual benefit.

The articles we have collected here for the first time take up the narrowly conceptual-analytical tasks introduced by social robotics. Thus robophilosophy as it is undertaken here is not so much reflective philosophy *of* social robotics, but rather constructive philosophy (and cognitive science) *for* social robotics. Keeping the analytical focus on sociality and normativity the chapters explore conceptual core questions arising with the very idea of an artificial social agent, such as the following: What does sociality mean in the context of social robotics? Can robots be social in the same sense as human beings are social? In which sense and to what extent will robots be able to simulate human sociality? In which sense of normative agency will robots be able to enter into normative relationships with human beings or with each other? What is the place of robotic entities in social ontology? Can robots be persons or normative agents? Can they become members of human societies?

To clarify these questions, one cannot resort to an "established and comprehensive" theory of human sociality, let alone of human-robot interactions, for there are none yet – interdisciplinary research on human sociality and human-robot interaction (HRI) both are young fields. Moreover, both research areas, though the second more than the first, struggle with the fact that we lack the theoretical vocabulary to articulate forms of sociality which deviate from the philosophical model case involving individual subjects with reciprocal capacities of intentionality and normative reasoning.[1]

[1] In a similar vein, the contributions to (Misselhorn, 2015) explore whether and how philosophical notions of collective agency and cooperation can be brought to bear on human interactions with artificial agents of all kinds. Since human interactions with virtual agents on-screen differ profoundly from human interactions with robots, i.e., artificial agency within the physical space of human social interaction, it is important, we believe, to differentiate the discussion. Moreover, we believe it useful to put the investigative focus first on the fact that the phenomena of human-robot

Chapter 2 thus introduces in outline a conceptual framework for the description of “asymmetric” – i.e., not fully reciprocal – forms of sociality. The framework for “Simulated Social Interactions” (SISI) which Johanna Seibt sketches here, can be used to formulate claims and investigatory hypotheses for empirical and conceptual studies of social interactions in general, and human-robot interactions in particular. She first draws attention to what she calls the “soft problem in the ontology of social robotics” – i.e., the task of describing human social interaction with robots without ascribing to robots beliefs and intentions from the outset. To address this problem, she argues, we need to treat the “as if” mentioned in many descriptions of human-robot interaction as the “as if” of simulation, not of fictional pretend-play. She distinguishes five kinds of simulation or partial realization, formally defined in terms of relationships between process systems. Then she shows how these five modes of simulation can be used to define for any type of two-agent interaction a matrix of “simulatory expansions” of that interaction type. Simulatory expansions of social interactions can be used to articulate different kinds and degrees of sociality. This allows us to clarify which conceptual and empirical reasons we traditionally draw on, and which we could draw on, when we determine the requirements for social interactions. It also allows us to identify precisely where, and on what grounds, we require different capacities and affordances in our interaction partners when it comes to social interaction with robots as opposed to, e.g., interactions with children or animals. In this way, she suggests, we can gradually, in a feedback loop of conceptual and empirical research, build up a taxonomy of social interactions, symmetric and asymmetric, for humans, robots, and other candidates for social interaction partners.

The subsequent chapters undertake this sort of research, combining terminological differentiations with substantive claims about sociality and normativity that are grounded in empirical research in cognitive science and psychology.

Chapter 3 contests the presuppositions of the traditional debate about sociality by a substantive account of sociality developed within a broader systematic context of a naturalist process metaphysics. Mark Bickhard, a protagonist of the “interactivist” paradigm in cognitive science and philosophy of cognition for many years, here shows how the basic questions about robot sociality are reconfigured once we proceed from the assumption that the world consists of processes rather than substances or other “static” types of entities. In a first step he recapitulates in broad strokes some central arguments of his earlier work, stressing that process ontology is in close contact with science at all scales and allows for a coherent account of emergence as dynamic organization of other processes. Some emergent processes are the dynamic organizations of systems far from thermodynamic equilibrium, i.e., self-maintaining and recursively self-maintaining systems, which, Bickhard argues, realize normative functionality in nature. On this thermodynamic account of normative function we can make sense of the idea that one process functionally presupposes another,

interaction challenge the classificatory concepts of social ontology at its most basic level, namely, the very notions of sociality and normativity themselves.

which, Bickhard argues, in turn can be used to develop an account of representation that is not troubled by "grounding" problems. In a second step, Bickhard sketches how elements of social reality (including language and personhood) arise in this naturalist framework. In order to solve coordination problems agents introduce "situation conventions," which are "emergent and implicit," that is, not based on any preceding reasoning. In a third step, Bickhard considers the possibility of robot sociality. The handling of situation conventions requires that the agent "has a stake in the world," i.e., interests that drive the establishment and repair of social relations. But "stakes," Bickhard claims, are tied to the architecture of biological far-from-equilibrium systems. Thus, Bickhard concludes, robots would be able to *be* social rather than merely simulate social behaviour if they were far-from-equilibrium systems with an architecture that enables emergent normativity akin to biological systems. While this may be possible, contemporary robots are not genuine social agents.

The chapter by David Eck and Alex Levine continues the theme of questioning longstanding assumptions in our understanding of sociality from the process-ontological perspective of the interactivist approach. Here the focus is on the traditional assumption that sociality presupposes individual subjectivity; from this assumption it would follow that robots, which are not individual subjects, cannot possibly engage in social relations. The authors begin by highlighting several paradoxical consequences of this assumption, which they call the "priority of individuality" thesis, in other areas of philosophy that share the following basic structural problem arising for traditional theories of sociality. As long as the explanandum, e.g., sociality, of the theory is conceived of as a relation, we need to postulate that there are different individuals between which such a relation can hold; but the relation in question is defined in such a way that the differences between these individuals are irrelevant. The structural problem of co-constitution of relatum and relation, or specifically, of individual and society – well familiar from the philosophy of German idealism – is best solved, the authors suggest, by turning to a purely process-based account, and specifically to the enactivist approach to cognition and sociality. With focus on the work by H. De Jaegher, the authors discuss recent proposals for which interactions must be added to largely autopoietic or self-maintaining processes to capture the characteristics of social coordination. Eck and Levine argue that these current process-based characterizations of sociality obscure, or even fail to accommodate, first-person experience and the historical dimension of subjectivity. To avoid that the process-based approach, which as such is on the right track, results in "hollow collectivism" it is important, the authors suggest in conclusion, to integrate data from qualitative research into the set of empirical data to be modelled.

The chapter by Jedediah Allen and Hande Ilgaz contributes an important piece of empirical research to the vision of a process-based theory of sociality suggested in the previous three chapters. The generic references to "social coordination" and "social relations" in Chaps. 3 and 4 raise the question of whether a process-based approach is strong enough also to describe the details of social learning, which in turn can be used to explain the complexity of social cognition and social behaviour in adults. Only if a purely process-based description of social learning is possible, one might argue, can one at all entertain the idea that the complex competences of sociality are not tied to a special kind of substance but could be implemented in a robotic system with the same process architecture. Allen and Ilgaz address this question by illustrating how one might explain social learning in general and imitation learning in particular from an interactivist or "action-based" perspective. They argue that imitation itself undergoes learning and development – selective imitation and over-imitation does not need to be understood as reflecting unique capacities of the human mind but can be treated as an instance of social meta-learning, i.e., of learning how to respond to the interaction possibilities of a learning situation that involves other agents. Such meta-learning capabilities will be required of social robots as well. On the interactivist account, an object is represented by its affordances for interaction; during imitation learning these affordance structures are modified but, this is the authors' decisive point, not only by the interaction with the object but by interaction with the entire situation in which the object is embedded and together with the affordances provided by other agents in the situation. In an empirical study Allen and Ilgaz show how this perspective on social meta-learning can be applied to an ongoing debate about why children imitate causally unnecessary actions while learning about a new artifact (i.e., over-imitate). The authors suggest that over-imitation in children shows that the affordance structures representing objects are acquired together with the social affordances of the situation and thus as "cultural artifacts."

The account of social cognition presented in the previous chapter does not presuppose the capacity of social agents to read each other's minds. The next chapter, by Vìctor Fernández Castro, endorses this view, this time not from the perspective of developmental psychology but from the discussion of social cognition in philosophy of mind, drawing on evolutionary biology. Fernández Castro first rehearses the traditional position of the "mindreading" approach, which presupposes that social agents can represent the other's mental states (beliefs, intentions, and emotions) and that this capacity is central for social behaviour that involves the prediction and explanation of the actions of others. In contrast, the "mindshaping" approach – developed by Mameli – holds that we can predict and explain the actions of others merely by assuming that others follow rational norms just as we do. Fernández Castro argues that the mindshaping approach is the superior model for human social cognition since – among a list of other strong arguments – unlike the mindreading approach it can explain both cases of cooperation without prediction and cases of prediction of actions. In a further step Fernández Castro identifies three areas where the mindshaping approach can be of use for social robotics. Due to the emphasis on developmental aspects mindshaping provides a testable model for "developmental robotics," which combines developmental psychology and social robotics research by implementing developmental models of social cognition. In addition, it can shed light on the relationship between general and social intelligence. Finally, mindshaping mechanisms may have direct implications for the design of companion robotics.

The next chapter can be read as an oblique demonstration of this last claim that the “mindshaping” approach can be used as a basic principle of software architecture. Here Felix Lindner and Carola Eschenbach develop a framework of affordances and affordance spaces that could be implemented and applied in order to enable social robots to spatially coordinate their actions in the presence of human beings and other robots. Nowhere in this framework are minds read or represented – the social behaviour is engendered by affordance spaces that reflect social norms. The authors note that although there have been several studies on affordances in robotics, most of them have focused on specific interaction contexts and paid too little attention to the effects that agents’ behaviours can have on each other’s affordances. Their focus is on spatial activities: They develop a framework in which various regions surrounding agents, objects, and, more generally, affordances can be defined in order for robots to be able to infer minimally obtrusive movement trajectories and locations to position themselves or other objects. One example would be avoiding overlaps between affordance spaces because they might indicate that action opportunities of other agents might become blocked. A careful analysis of affordance spaces during planning, for instance, enables robots to display behaviour that is more readily perceived as socially acceptable as the authors demonstrate via several examples and case studies.

The next three chapters investigate specific conceptual issues for human-robot interaction: Can humans and robot perform a joint action? Can they have joint commitment? And finally, which, if any, “understanding” of norms can humans and robots have in common?

In their contribution Aurélie Clodic, Elisabeth Pacherie, Rachid Alami, and Raja Chatila study the general conditions for joint action between humans and robots. They note that human joint action has been intensively studied within psychology and philosophy, and try to see to what extent the results of those studies can be connected to the field of human-robot interaction. The authors find parallels, in the single-agent case, between philosophical accounts of intentional action and layered architectures proposed in artificial intelligence and robotics. Such connections motivate them to look for similarities also in the case of joint action. They analyse several requirements and coordination processes that have been discussed in the case of human joint action and translate them to the case of human-robot joint action. Their study culminates in taking the first steps toward what they call a “framework for joint action” that describes three layers at which human-robot joint action can be represented. Such a framework can be used not only for analysing information needs and representational processes involved in joint action but potentially also for implementing robots capable of joint action with humans.

Alessandro Salice and John Michael focus on the notion of commitment which according to them is central in human social interaction including joint action. They will study the possibility of robots making commitments to humans and humans being motivated to honour commitments made to robots. They argue that even though full-blown commitments might be impossible in interactions between humans and robots, it might be useful to aim at designing robots that stimulate in humans tendency to act as if commitments were effective. This might suffice to acquire the potential social benefits that commitments generate such as predictability of action and higher level of engagement. Salice and Michael present several challenges that designing such robots might involve and consider ways to address them. In addition to studying ordinary interpersonal commitments they devote special attention to joint commitments that are associated with joint action, and consider whether humans could form groups with robots and whether psychological processes of group-identification could be triggered in cases of such mixed groups.

Johannes Brandl and Frank Esken investigate in which sense we can justifiedly say that robots are "social" agents, which they take to turn on the question in which sense we can justifiedly say that robots "understand norms." In interaction with empirical research Brandl and Esken develop three criteria which in combination characterize full-scale normative competence. Each one of these criteria defines a "stage" of social behaviour or "social intelligence," they suggest. At the first stage, agents conform to regularities that have a "social foundation" in the sense that they are behaviours of a group that are enforced by the dominant members of the group. This form of social behaviour, which is motivated by the individual's instrumental rationality, is also displayed by animals, e.g., by "rules" for the treatment of infants among chimpanzees. The second stage of social behaviour is reached when agents follow a rule because others expect it from them. When young children at the age of three protest against a violation of a norm they exhibit this level of understanding of norms or rules, Brandl and Esken argue. Here agents not only behave in accordance with a rule for instrumental reasons but they follow the rule *as* a rule, as a matter of "social practical rationality." The normative force of the rule here rests on social expectations. Finally, at stage three agents follow a rule with the understanding that the rule applies in all cases of this kind, i.e., the normative force of the rule lies in the rule itself. Only social behaviour at this last stage, Brandl and Esken suggest, involves a genuine understanding of norms as norms, and only at this last stage we can justifiedly speak of social agency in the full sense. To the extent that robots are unlikely to be able to operate on the third stage of social intelligence, we should refrain from considering them as social agents, even though they may exhibit social behaviour at the first two stages.

The final two chapters again broaden the scope of inquiry and consider the questions what we will need to become when we begin to share our interaction space with robots, how we will need to invent new venues to preserve our capacity to care, i.e., to act freely for the sake of values, and what robots would need to become if we were to integrate them into our life-form.

In the penultimate chapter, Antonio Carnevale approaches the question whether robots can figure as social or normative agents from the perspective of foundational reflections engaged by recent philosophy of technology. Modern electronic technology uses software and thus representations at levels where older technology merely employs causal translations; thus modern technology introduces new levels of abstractions or declarative representations. With care robots the abstractive declarations of technology enter into the most intimate region of human interaction space. These dimensions we need to have in mind, Carnevale argues, when we ask what kind of care relationship could exist between a robot and a human. The rational response, he suggests, is to reform certain concepts for social relations in such a way that interactions with technology are included from the outset. As long as we merely insist that care must be based on love or solidarity, we have only the option of deploring that the increasing use of “carebots” in the healthcare sector will decrease the “humanity” in caretaking. But we cannot give up caring, which is, in Carnevale’s view, an essential aspect of a human life. Caring means setting values and choosing freely to pursue them. The dilemma: “no care in healthcare versus no technology in healthcare” can be avoided, Carnevale argues, once we recognize that there are other ways to care – we care when we are responsible. Thus the responsible integration of technology into healthcare and other social interaction spaces can become our new form of caring. Carnevale sketches three components that characterize such a responsible way of policy making that he calls the “I Tech Care” approach.

In the last chapter, Hans Bernhard Schmid considers the possibility of robots’ participating in our “life-form” (i.e., our lifeform as macrobic organisms and our Wittgensteinian form of life as complex of social practices) and of them developing a life-form of their own. The chapter’s main aim is to critique the notion of a robot by an analysis of several various misplaced dichotomies in the characterization of robots which, he suggests, hamper the current discussion about robots in society. Schmid begins with a closer look at the specific place where the term “robot” enters human cultural history, Karel Capek’s play “Rossum’s Universal Robots.” Using an interpretation of Capek’s play as an expository foil throughout, Schmid first argues against the distinction between routine work and discursive practices that underlies many “robo-sceptical” views. He draws on the distinction between “mimeomorphic” and “polymorphic” actions, i.e., actions with implicit and explicit complexity, and argues that it is by no means clear that robots cannot perform actions that simulate actions with implicit complexity, as these are involved in our social and discursive practices. The second dichotomy he questions is the division between natural life and mechanical artifacts. Finally, reflecting on the ambiguity of the end of Capek’s play he draws attention to the close connections between our notion of life and the valuative dimension – the specific values that we aim to realize in a human life seem to be tied up with the classification of something as a life-form. While the play can afford an aporetic ending on the question of what robots need in order to be able to participate in our life-form, we are in a situation, Schmid points out, where we cannot just wait and see how things play out.

The common theme running through all the papers is an attempt to understand conditions of sociality and normativity in interactions between humans and robots, and the conceptual problems that result in trying to fit robots in the conceptual and normative frameworks that have evolved for understanding human sociality and interaction. We hope that the book will aid readers to gain a better understanding of the descriptive issues involved, and thereby acquire new perspectives on the ethical questions raised by social robotics.

References

Coeckelbergh, M. (2012). *Growing moral relations: Critique of moral status ascription*. Houndmills/Basingstoke: Palgrave Macmillan.

euRobotics aisbl. (2013). Strategic research agenda for robotics in Europe 2014–2020. Retrieved from http://www.eurobotics-project.eu. Accessed 14 Sep 2016.

Gunkel, D. (2012). *The machine question*. Cambridge, MA: MIT Press.

Misselhorn, C. (Ed.). (2015). *Collective agency and cooperation in natural and artificial systems: Explanation, implementation and simulation* (Vol. 122). Cham: Springer.

Nørskov, M. (Ed.). (2015). *Social robots: Boundaries, potential, challenge*. Farnham, Surrey: Ashgate.

Seibt, J. (2017). Robophilosophy. In R. Braidotti & M. Hlavajova (Eds.), *Posthuman glossary*. London: Bloomsbury.

Seibt, J., Hakli, R., & Nørskov, M. (Eds.). (2014). *Sociable robots and the future of social relations: Proceedings of Robo-philosophy 2014* (Vol. 273). Amsterdam: IOS Press.

Seibt, J., Hakli, R., & Nørskov, M. (Eds.). (2018, forthcoming). *Robophilosophy: Philosophy of, for, and by social robotics*. Cambridge, MA: MIT Press.

Sharkey, A., & Sharkey, N. (2012). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology, 14*(1), 27–40.

Sullins, J. P. (2005). Ethics and artificial life: From modeling to moral agents. *Ethics and Information Technology, 7*(3), 139–148.

Sullins, J. P. (2006). When is a robot a moral agent? *IRIE: International Review of Information Ethics, 6*, 23–30.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Chapter 2
Towards an Ontology of Simulated Social Interaction: Varieties of the “As If” for Robots and Humans

Johanna Seibt

Abstract The paper develops a general conceptual framework for the ontological classification of human-robot interaction. After arguing against fictionalist interpretations of human-robot interactions, I present five notions of simulation or partial realization, formally defined in terms of relationships between process systems (*approximating*, *displaying*, *mimicking*, *imitating*, and *replicating*). Since each of the n criterial processes for a type of two-agent interaction $\mathfrak{I}$ can be realized in at least six modes (full realization plus five modes of simulation), we receive a $(6^n \times n) \times (6^n \times n)$ matrix of symmetric and asymmetric modes of realizing $\mathfrak{I}$, called the “simulatory expansion” of interaction type $\mathfrak{I}$. Simulatory expansions of social interactions can be used to map out different kinds and degrees of sociality in human-human and human-robot interaction, relative to current notions of sociality in philosophy, anthropology, and linguistics. The classificatory framework developed (SISI) thus represents the field of possible simulated social interactions. SISI can be used to clarify which conceptual and empirical grounds we can draw on in order to evaluate capacities and affordances of robots for social interaction, and it provides the conceptual means to build up a taxonomy of human-robot interaction.

Keywords Social ontology • Classification • Social robots • Sociality • Simulated interaction • Taxonomy • Human-robot interaction

J. Seibt (✉)
Research Unit for Robophilosophy, Department for Philosophy and History of Ideas,
School of Culture and Society, Aarhus University, Denmark Jens Chr. Skous Vej 7,
DK-8000, Aarhus C, Denmark
e-mail: filseibt@cas.au.dk

R. Hakli, J. Seibt (eds.), *Sociality and Normativity for Robots*, Studies in the
Philosophy of Sociality 9, DOI 10.1007/978-3-319-53133-5_2

2.1 Introduction

At first blush, the notion of a "social robot" does not seem to make much sense. If we use prototype semantics to interpret the terms involved, we end up with a category mistake—the currently prototypical robot, i.e., the industrial robot performing repetitive movements irrespective of context, is not the kind of entity that could engage in prototypical social interactions such as reaching an agreement or meeting for a cup of coffee. If we set prototype semantics aside and turn to explicit definitions, matters do not seem to improve at all. For, on the one hand, researchers in "social robotics" advise us to abandon all hope with regard to conceptual fixations within a fast changing field—"the concept of robot is a moving target, we constantly reinvent what we consider to be 'robot'" (Dautenhahn, 2014). On the other hand, while the term "social" still is taken to be amenable of definition, so far there is only a research path *towards* a comprehensive account of human sociality but not yet a theory with stabilized terminology. Research on social ontology in philosophy is yet to come into full contact with efforts by anthropologists, linguists, psychologists, and cognitive scientists to create an integrated study of human sociality (Enfield & Levinson, 2006). In short, we have only begun to determine, based on conceptual and empirical research, what could or should be understood by the qualifier "social," and the class of items we wish to qualify with this yet rather vague predicate—the reference class of "robots"—is under construction, even constitutively perhaps.

This is an unusual situation. The philosopher's uneasiness can only increase, however, upon reading that a "social robot" is designed to enable people to "interact with it as if it were a person, and ultimately as a friend" (Breazeal, 2002, p. xi). This one-liner nicely encapsulates three assumptions that, from the philosophical point of view, raise conceptual issues of breath-taking significance.

The first issue concerns the idea that human-robot interactions could qualify as instances of *social* interactions—are we justified in applying the term "social interaction" even when the relevant capacities for sociality are not distributed symmetrically over the interacting systems? In short, could there be, in principle, *asymmetric* social interactions?

The second issue centers on the type of capacities that robotic systems could possibly exhibit. Given the current state of robotics, most of those capacities that we consider distinctively human, such as linguistic communication and intelligent behavior, are merely approximated by robotic systems. But can capacities for sociality at all be *approximated*? Which degrees and kinds of approximation are admissible, and can robots have the relevant approximative capacities?

The third issue concerns differences in the normative dimension of classificatory predicates. The predicates (i) *x is a y* and (ii) *x can be interacted with as if it were some y* carry different semantic restrictions and signal different practical implications. Will the practice of human-robot interaction change the semantics and normative significance of ascriptive predicates (e.g., *x is a person*), i.e., will we come to allow for anything normatively to count as a person if our interactions with it resemble interactions with a person?

The philosophical debate about social robotics so far has concentrated on ethical and moral aspects, either sidestepping conceptual issues or addressing them obliquely within a specific context of application. This focus made good sense, given that the basic design goal of social robotics – artifacts that engage humans in interaction types that we used to reserve for humans only – generates a host of urgent questions of policy and legislation. Up until recently philosophy of social robotics mainly engaged in "roboethics" (Veruggio, 2006) understood as an area of applied ethics that discusses benefits and disvalues of social robotics applications in elderly care, schools, or therapeutic contexts (cf. e.g. Sullins, 2008; Sharkey & Sharkey, 2012; Sparrow & Sparrow, 2006; Coeckelbergh, 2010; Vallor, 2011).

Increasingly, however, the scope of roboethics has widened to take up general methodological issues, exploring the challenges of social robotics to traditional paradigms of moral agency and moral patiency (cf. e.g. Gunkel, 2012; Coeckelbergh, 2012). As these investigations bring out clearly, supported by empirical research on human-robot interaction, the phenomenon of human-robot interaction contains conceptual ambiguities that are directly linked to ethical judgments (cf. e.g. Kahn, Freier, Friedman, Severson, & Feldman, 2004).

We need to conceptually clarify the phenomenon of human-robot interactions in all its diversity, in order to make progress in the professional and public ethical debate about social robots. But which methods should we apply if the traditional essential differences between *res extensa* versus *res cogitans* are dissolving? Postphenomenological philosophy of technology recommends the radical step of abandoning ontology. However, there are good epistemological reasons, in my view, why roboethics should rely on both empirical research and ontology to provide a fine-grained analysis of human-robot interactions. Since public policy and legislation are guided by ethical recommendations, it is problematic if ethical recommendations are based on subjective impressions about what a robot "is" or "is not"—it may be presumptuous, for example, to suggest without more detailed ontological and empirical investigations that human interactions with carebots are "too undignified for a human to enter into" since carebots are "totally inadequate surrogate products" (Danish Ethical Council, 2010). There is increasing empirical evidence that in certain contexts of therapy and caretaking (e.g., autism and dementia) humans benefit more from the interaction with robots than from human interaction (Cabibihan, Javed, Ang Jr, & Aljunied, 2013; Leyzberg, Avrunin, Liu, & Scassellati, 2011).

More generally speaking, the Western democratic conception of political authority is grounded in the epistemology of discourse-theoretic validation; ethical judgements are to be warranted by "rational discourse," which in turn presupposes normative classifications that are established by intersubjective methods of empirical and theoretical inquiry. When postphenomenologists dismiss "ontology" in favor of a postphenomenological or "relational" approach focused on subjective "takings-as," they miss out on an important ally that can translate the "relational approach" into the conceptual format needed for the sort of rational discourse that in Western societies still counts as epistemic warrant for ethical and political decision making. While—correctly, in my view—ridding themselves of the murky bathwater of traditional realist "substance–attribute" ontology, postphenomenologists inadvertently throw out the baby of pragmatist analytical ontology, which alone is currently capable of developing sufficiently differentiated and intersubjectively justifiable normative classifications of interactions.[1]

The following considerations, written from the perspective of pragmatist analytical ontology, are intended to offer clarifications and classificatory tools that should prove useful for the ethical debate about social robotics applications. In addition, the conceptual distinctions offered below may also have heuristic value for the design of empirical studies of human-robot interactions, especially when undertaken in the context of research on social cognition. I will outline a conceptual framework on the basis of which any extant and future type of human-robot (and human-human) interaction can be classified, so that a comprehensive and detailed taxonomy of human-robot interactions can be built up incrementally.

I begin in Sect. 2.2 with a brief reflection on the predication contexts "*x* treats *y* as *z*," "*x* interacts with *y* as if it were *z*," and "*x* takes *y* as *z*." As I explain, these contexts not only govern different modes of constructionality (make-believe, fictionality, conventionality or "social reality") they also behave differently with respect to precisely those ascriptive terms that social roboticists are wont to use in their descriptions of human-robot interactions. I argue that for conceptual reasons we cannot adopt the—temptingly easy—strategy of treating human-robot interactions as fictionalist analogues to human-human interactions. Thus we are saddled with what I call the "soft problem in the ontology of social robotics," namely, the problem of how to describe human-robot interactions, from a second-person and third-person point of view, given that our concepts for human social interactions as such are inapplicable. I suggest addressing the soft problem by developing a theory of simulated social interaction. In Sect. 2.3 I set out on this task and define five notions of simulation, formulated as relationships between process systems: replicating, imitating, mimicking, displaying, and approximating. In Sect. 2.4 I sketch how these distinctions can be used to define for each interaction concept $\mathfrak{I}$ a "simulatory expansion" of $\mathfrak{I}$ and show how the latter can be used to characterize human-robot interactions. In Sect. 2.5 I address the question of which sorts of simulations of a social interaction $\mathfrak{I}$ can qualify as a social interaction. Given that we currently have not one but several competing notions of sociality, I suggest that we should abandon the idea of a dualist distinction between social and non-social interactions; rather, we should conceive of sociality as a matter of degree. I show how we can use the notion of a simulatory expansion of a social interaction $\mathfrak{I}$ to create a taxonomy for human-robot interactions. As I shall explain, this taxonomy promises to enable us to address two tasks in one go. On the one hand, it should be possible to integrate within this taxonomy competing accounts of sociality with more and less restrictive requirements. On the other hand, with the descriptive tools of the taxonomy we

[1] Note that pragmatist analytical ontology is not committed to the facticity or even the possibility of rational discourse, just to its utility as regulative idea and regulated praxis.

can create a comprehensive conceptualization of simulated sociality that allows for differentiated descriptions of human-robot interactions from a second-person and third-person point of view.

2.2 Against Fictionalist Deflations of Robot Sociality

Social robotics creates a practical space that we currently have great difficulty conceptualizing. While adults interacting with social robots display signs of conceptual disorientation, children tend to resolve the conflict pragmatically by using contextual categorizations of robots ("alive enough to be *x*"), or hybrid categorizations ("mechanical yet with feelings," cf. Bernstein & Crowley, 2008), or practically relate to robots in ways that conflict with the chosen categorization (cf. Clark, 2008; Turkle, 2011). While most researchers might agree that we need a new ontological category (cf. Kahn, Friedman, Perez-Granados, & Freier, 2004) in order to conceptualize human interactions with "social robots" so-called, it is currently far from clear how we should go about this task. Do we need a new concept for the item we are interacting with, the robot, as a new sort of artificial "living thing"? Or, leaving the robot in the familiar category of machines, do we need a new concept for our responsive dispositions for interacting with such machines, e.g., as dispositions for machine sociality? Or should both relata, robot and human, retain their traditional conceptualizations and should we introduce a new sort of relation between them: fictional social relations?

These are three basic strategies that traditional substance–attribute ontology has to offer, as well as various combinations thereof. Once we abandon the traditional substance paradigm, however, new systematic options for categories come into view. In this paper I will suggest quite a different line of approach, however. It is not *one or two* new ontological categories that we need in order to conceptualize human-robot interactions, I shall argue, but an entire new classificatory framework for simulated social interactions. In order to motivate this approach I will begin with some general remarks on the idea that human-robot interactions could be *fictional* social interactions.

Humans interact with their environment not only (i) physically but also (ii) symbolically, i.e., representing physical features, and (iii) "figuratively," i.e., assigning to representations of physical features new interactive significances. Consider the following three types of "figurative" interactions of a human with her or his environment (an object, another human, an event etc.),[2] which I introduce here together with their characteristic linguistic forms, i.e., linguistic expressions that are typically used in descriptions of the participant, who experiences the interaction

[2] For the sake of simplification I shall throughout this paper assume that an interaction has just two participants (i.e., two human participants, or a human and a robot).

from a second-person point of view, and descriptions of an observer from a third person point of view, respectively.

1. *Make-believe or pretend-play*. The participant describes the object she is interacting with according to the interpretational rules of the make-believe scenario (e.g., "I tied my horse to a branch"), and the observer describes the interaction as make-believe scenario extrapolating from normal interpretational rules based on similiarity (e.g., "she treated the stick *as* a horse").
2. *Fictional interaction*. The participant describes the interaction in accordance with the conventions of the fiction (e.g., "I greeted the king and was admitted to the court"), the observer describes the participant's behavior together with its significance relative to the conventions of the fiction (e.g., "she bowed as if she were greeting a king and he moved his head and right hand as if he were admitting her to approach").
3. *Socially instituted interaction*. The participant describes the interaction in accordance with extant social conventions (e.g., "I showed him the receipt") and the observer describes the participant's behavior and its social significance (e.g., "she handed him a piece of paper that counts as proof of payment").

Observer descriptions of make-believe scenarios are linguistically typically signaled by the phrase "*x* treats (considers) *y* as *z* (or: as if it were *z*)," while fictional interactions are typically described by "*x* interacts with *y* as if it were doing *z*," and socially instituted interactions are typically characterized by the phrase "*x* takes (object or interaction) *y* to count as *z*."[3]

With these distinctions in mind, let us now consider the following formulations of the design goals of social robotics and descriptions of human-robot interactions, respectively:

1. "We interact with [a sociable robot] as if it were a person, and ultimately as a friend" (Breazeal, 2002, p. ix).
2. "Ideally, people will treat Kismet as if it were a socially aware creature with thoughts, intents, desires, and feelings. Believability is the goal. Realism is not necessary" (Breazeal, 2002, p. 52).
3. "This also promotes natural interactions with the robot, making it easier for them to engage the robot as if it were a very young child or adored pet" (Breazeal, 2002, p. 100).
4. "I find people willing to seriously consider robots not only as pets but as potential friends, confidants, and even romantic partners" (Turkle, 2011, p. 26).
5. "... social robots—the class of robots that people anthropomorphize in order to interact with them" (Breazeal, 2003, p. 167).

Formulations (2) and (4) describe human-robot interactions as make-believe scenarios where an *object* is treated as something else, or as if it were something

[3] For the sake of the argument in this section I operate here with a simplified version of Searle's definition of social reality: "For all kinds *Z*, instances of kind *Z* is part of social reality iff there are *X*, *Y*, and *C*: *X* takes *Y* to count as a *Z* in circumstances *C*," (cf. Searle, 2010).

else. Humans mainly engage in make-believe projections during childhood (pretend play), but many people allow themselves to entertain the special type of *anthropomorphistic* make-believe projections also in adult life, as a form of conscious, self-ironic sentimentality. Formulation (5) explicitly clarifies that the design of social robotics consciously targets our capacities and dispositions for anthropomorphizing make-believe projections.

Very little indeed is needed to anchor such projections and to allow us to treat an object as a human being or as a companion. You may treat the tree in front of your window as the master of the garden just because it overshadows all other plants of the garden, and you may treat your car as your companion or adversary just because its start-up performance in difficult weather resembles reliable or malicious actions.[4]

Projections of make-believe ("treating *x* as if it were *y*") are based on physical or functional similarities or analogies between, on the one hand, features or doings of an intentional agent, and, on the other hand, static or dynamic features of a natural or artificial item (Walton, 1990). Fictional interactions are also based on similarities or analogies, but there is an important difference between "treating some *x* as if it were to do *y*" and "interacting with *x* as if it were *y*." When I treat my car as companion and greet it, the car does not perform any distinctive behavior in return; in contrast, a NAO-robot reacts to a human greeting with greeting behavior, or even autonomously displays greeting behavior to elicit a human greeting. In other words, make-believe scenarios typically are one-sided analogical projections where only the human agent involved executes the actions in the template of actions and reactions that define the interaction. In contrast, in fictional interactions both agents[5] behave in ways that resemble the actions and reactions prescribed by the interaction template. Typically we connect fictional interactions with role-play, where the relevant fictional conventions are understood by all agents involved. But a fictional interaction can also be said to take place even if one of the agents is not conscious or is not aware of any convention of fictionality being in place. For example, as long as the behavior of my dog resembles a greeting or as response to my greeting, one can say that I interacted with him as if we were greeting each other; and as long as the behavior of a hypochondriac resembles pain relief after medical treatment, one can say that by administering a placebo I interacted with him as if he were in pain. In short, whether a scenario is make-believe or a fictional interaction

[4] Here I bracket the question whether "anthropomorphizing" is the right label for make-believe projections of this kind. Treating something as companion or foe does not necessarily imply treating it as human being. Especially if one applies the "non-exceptionalist" notions of sociality I discuss below, one might argue that even though human beings are the primary instances of social actors, our long-standing practice of projecting social roles onto natural things and artifacts is a way to *"socialize"* the world, not to "anthropomorphize" it.

[5] Throughout this paper I use the term "agent" in the broad disjunctive sense where it refers either to agents proper, i.e., conscious living things that can act on intentions, *or* to inanimate or living items that are causal origins of events that resemble (and thus can be treated as) actions, i.e., the doings of agents proper.

depends on whether the agents engender occurrences that resemble the actions and reactions of an interaction template.

Applying these considerations to quotations (1) through (5) above, human-robot interactions are described as make-believe in (2), (4) and (5) (here explicitly), and as fictional interaction in (1) and (3) (in the context of this passage Breazeal explains that her robot Kismet is programmed to produce behavioral patterns that make it more "natural" for people to interact with it "as if it were a very young child or adored pet"). None of the quoted announcements of the design goals of social robotics use formulations that are the characteristic indicators of social actions, i.e., the movements of the robot are not supposed to be "taken as" or to "count as" certain actions, nor are the participating humans said to exhibit behavior that "counts as" an action with social significance. In other words, the quoted passages do not describe human interactions with robots as scenarios where a social action *de facto* occurs.

In comparison with other descriptions of the design goals of social robotics in terms of intentionalist vocabulary, where robots are said to "perceive" their environment or even are said to be "aware" of it (cf. Fong, Nourbakhsh, & Dautenhahn, 2003, p. 145), the strategy of using linguistic forms that express the fictional irrealis of sociality rather than the realis may seem a very useful device of professional caution. But upon a closer look any attempt to deflate the question of robot sociality by using social vocabulary in fictionalist embeddings are bound to be unsuccessful, as I shall argue now.

Consider again formulation (1). Could we ever interact with anything *"as if it were a person"*? As just mentioned, both make-believe and fictional interaction are based on resemblances or analogies between descriptive aspects of entities and interactions. But the predicate "person" is not a descriptive predicate. When we call an entity a person, we thereby make certain commitments in the performance of that very utterance—we are not describing features but announce that certain commitments are undertaken. The performative-ascriptive use of language is not limited to promises and explicit declarations—it pervades our vocabulary for social interactions. Importantly, performative-ascriptive predicates cannot be embedded in contexts with fictionality markers. One cannot perform a linguistic-pragmatically and conceptually coherent speech act by uttering "It is as if I hereby promise you ...," nor "what I will say now is a bit like a promise ... " Similarly, if we treat some x as a person we are committed to taking x to count as a person—that is, assuming that we wish to abide by linguistic norms and the actions they entail, we *must* interact with x *as* a person.

This is due to two facts of social reality.[6] The first fact is that commitments are strictly "bivalent"—they are either undertaken or they are not undertaken; pretending to undertake a commitment is simply *to fail* to undertake it. But if we cannot make fictional commitments, we cannot make fictional promises, nor can we

[6] In the context of this paper I take it that these two facts are self-evident elements of the "logic" of social practices; a more detailed discussion of the semantics of fictional discourse in application to the performative-ascriptive predicates for social and moral roles is in preparation.