To appear in: Braidotti, R. / Hlavajova, M. *Posthuman Glossary*, forthcoming.

Posthuman Glossary, entry "Robophilosophy"

J. Seibt

The term 'robophilosophy' stands for a fundamental systematic reconfiguration of philosophy in the face of artificial social agency. Unlike other systematic research initiatives in philosophy, robophilosophy is time-sensitive, directly motivated by technological developments, and proactive. Robophilosophy is a response to (1) projections of the explosive development of the robotics market in the third decade of the 21st century, and (2) to empirical evidence that the large-scale use of artificial "social" agents in public and private spaces of human social interactions quite likely will lead to profound disruptions of economic, social, and cultural practices in industrialized societies West and East.

The term 'robophilosophy' has wider currency in academic contexts since the inauguration of the bi-annual *Robophilosophy Conference Series* in 2014.[1] The term was coined by the author in 2013, in resonance with G. Veruggio's (2004) call for 'robo-ethics', in order to signal that the challenges of 'social robotics' go beyond ethical concerns and address all disciplines of philosophical research. Moreover, robophilosophy is a complex reconfiguration that engages three research perspectives at once—it is "philosophy *of*, *for*, and *by* social robotics" (Seibt et al, 2017). The following paragraphs will describe each of these three perspectives in greater detail; however, as also shall become clear in the course of the exposition, these perspectives form systematically connected trajectories and contributions to robophilosophy—here associated for illustration with one perspective—should more properly to be characterized in terms of locations within a three-dimensional research space.

The first dimension, philosophy *of* social robotics, takes the reflective stance of traditional philosophical research and investigates the conceptual implications of the phenomena of human interactions with robots that act in accordance with social norms. After a decade of empirical research in 'human-robot interaction studies' (HRI) there is sufficient evidence to show that humans accept robots as social interaction partners and even attribute to them moral standing. Given that these human reactions are sincere, they are counterevidence to (a) the Cartesian paradigm of subjectivity according to which self-consciousness, freedom, intentionality, normative agency, and epistemic and moral autonomy are a package deal, and (b) to traditional and still dominant philosophical conceptions of sociality that restrict the capacity for sociality to Cartesian subjects, or else postulate, with Hegel, constitutive mutual dependencies between the capacity of sociality and the capacities associated with the traditional model of subjectivity. Since the latter figures centrally in the legitimization of moral and political authority in Western democracies, there may be far-reaching repercussions of a pervasive practical reconfiguration of the relevant capacities (e.g., sociality without self-consciousness, normative agency of great economic power without freedom). In short, robophilosophy as 'philosophy *of* social robotics' tries to come to terms with the fact that the empirical evidence collected in HRI research goes

[1] See www.robo-philosophy .org. The notion of robophilosophy as expounded here summarizes general insights from collaborative research in the 'PENSOR group' (Philosophical Enquiries into Social Robotics, www.projects.au/pensor), with special acknowledgements to M. Nørskov, R. Hakli, R. Rodogno, S. Larsen, C. Hasse. , J. C. Bjerring, M. Damholdt, C. Vestergård, and R. Yamazaki. The PENSOR group (which has 10 core members and 6 affiliated researchers) is the first research group in Europe investigating philosophical aspects of social robotics with wide interdisciplinary scope, combining research competences in philosophy (ontology, philosophy of science, epistemology, logic, intercultural philosophy, ethics, political philosophy), robotics, anthropology, psychology, cognitive science, education science, and computer science.

against a built-in feature of Western thought that only humans are the kind of entity that can stand in social relations, and/or that standing in social relations confers upon humans exceptional capacities, as well as the rights and statuses adhering to these. Turkle's felicitous observation that "we live the robotic moment not because we have companionate robots in our lives but because the way we contemplate them on the horizon says much about who we are and who we are willing to become" (2011: 26) addresses the 'robotic moment' from anthropological perspective as a turning point in contemporary culture, while robophilosophy as philosophy of social robotics puts the 'robotic moment' into the wider perspectives of human socio-cultural and political history and explores its metaphilosophical implications as a game changer for philosophical research. Outstanding examples of philosophical interactions with social robotics that explicitly engage the metaphilosophical dimension are Coeckelbergh 2012 and Gunkel 2012, who relate the new ethical tasks arising with social robotics to the deconstructions of modern subjectivity that 20th century philosophy developed on purely theoretical grounds.

Another important task for 'philosophy *of* social robotics', the reflective dimension of robophilosophy, is to situate the phenomena of human-robot interactions within the larger context of philosophy of technology. As Nørskov 2015 observes, Don Ihde's phenomenological classification of 'human-technology relationships' must be fundamentally reworked to capture the peculiar complexities of the phenomenology of human-robot interactions. Interestingly, since robots are produced in high-technology societies West and East, philosophical reflections on social robotics quite naturally lead from auto-cultural hermeneutics into into cross-cultural comparative and intercultural philosophy of technology (cf. Nagenborg 2007, Funk et al 2009, Nørskov 2011, Nakada 2013).

The second dimension of robophilosophy, 'philosophy *for* social robotics,' employs standard methods of philosophical research such as conceptual analysis, method analysis, capacity analysis, phenomenological analysis, formal theory construction, and value-theoretic discussion for the sake of addressing theoretical problems in the research methodology of social robotics, and in order to guide the development of social robotics applications.

To begin with the foremost task of a philosophy *for* social robotics, roboticists and researchers in HRI (Human-Robot Interaction Studies) currently operate in an interdisciplinary domain (in the intersection of robotics, psychology, anthropology, and sociology) that suffers from the lack of a joint descriptive framework relative to which robotic capacities, human reactions, and human-robot interactions can be characterized in clear and precise terminology. Despite some early efforts to clarify and classify varieties of "social" robots (Breazeal 2003, Fong et al 2003), most researchers in social robotics use the epithet "social" without apparently being aware of the semantic commitments incurred by our current conceptual norms that govern the meaning of this term. As philosophical reconstructions of these conceptual norms make clear, however, we cannot simply relax the requirements for sociality in general without thereby effecting central regions of our inferential space (Hakli 2014). Rather, we need to consider sociality as a gradient notion and develop precise, differentiated descriptions of human-robot interactions that justifiedly can be said to realize various degrees and types of sociality. Currently robotic capacities are described metaphorically, using the intentionalist vocabulary of human actions and social interactions—robots are said to "answer," "recognize," "deliver," "respond," "collaborate," "smile," "greet," etc. At best such intentionalist idioms are bracketed by the 'de-realization operator' *as-if*: "We interact with [a social robot] as if it were a person, and ultimately as a

friend" (Breazeal 2002: ix). Here and elsewhere the preposition *as-if* is presented as the 'as-if' of fictionality and pretend-play, which has motivated ethical criticism of social robotics as engaging humans in inauthentic social relations. However, fictionalist interpretations of the sociality in human-robot interactions are incoherent; social relations cannot be 'fictionalized'—I cannot treat an item *as if* it were a person since the performance of such a social action is constitutive for its realization (Seibt 2014, 2016). Rather, the de-realization in question should be understood as the *as-if* of simulation, where simulation is a similarity relation on processes; the latter can be used fairly straightforwardly for the definition of a fine-grained classificatory framework for simulated social interactions and associated degrees and types of sociality allowing for asymmetric (non-reciprocal) distributions of capacities among interaction partners (ibid.). This switch from the 'as-if' of fictionality to the 'as-if' of simulation—which fundamentally changes the premises for an ethical evaluation of human-robot interaction—is the cornerstone for a comprehensive descriptive framework for the interdisciplinary field of HRI.[2]

The second task area of a philosophy *for* social robotics is to analyze in detail specific human capacities and social roles. For example, which kinds of functionalities would a robot need to have to able to provide "care" or to "teach" or to "coach"—in the sense relevant in, e.g., healthcare, language training, or dietary assistance, respectively (Vallor 2011, v. Wynsberghe 2015)? If robots are to be "friends" or "companions," which behavioral routines would they need to exhibit to be perceived as such (Sullins 2008)? These investigations are direct extensions of familiar capacities analyses in AI of human cognitive predicates; however, while the question whether computers really can "think" or "form new concepts" is mainly of theoretical interest, conceptual and phenomenological analyses of capacity requirements for social actions and roles immediately lead to ethical issues. This also holds for the capacity of ethical reasoning itself—investigations about *how* to implement ethical reasoning in machines—e.g., in military robots—are tied to the question of *whether* to do it and thereby relinquish control (Wallach 2010). In tandem with developing a fine-grained classificatory framework for the description of human-robot interactions, philosophy *for* social robotics thus must define a differentiated array of new notions of moral and legal responsibility for collective agency constellations that involve robots.

The third dimension of robophilosophy, philosophy *by* social robotics, represents a far-reaching methodological reorientation of philosophical research. As mentioned above, HRI research is an interdisciplinary field operating with quantitative, experimental, and qualitative empirical research. If philosophy becomes, as philosophy *for* social robotics, an integral part of HRI—as it must, due to ethical concerns—the standard philosophical methodologies (conceptual and phenomenological analysis, rational value discourse etc.) lose the relative autonomy that is traditionally credited to them. The research results of HRI not only force philosophers to rework traditional conceptions of normative agency, sociality, moral status, responsibility, etc., they also open up new ways of conducting 'experimental philosophy.' For example, by

[2] In other words, human-robot interaction is not 'a human playfully pretending to perform a social action towards a robot,' but 'a robot simulating the performance of a social action towards a human.' This does not betoken, however, that investigations of the 'as-if' of fictionality is irrelevant for HRI. Larsen (2016) shows that the contrastive comparison between discourse about properties of fictional characters and discourse about robotic capacities is of important heuristic value for the semantic regimentation of descriptions of human-robot interactions formulated with the derealization operator 'as-if.'

implementing ethical reasoning in robots philosophers can investigate by construction and experiment which, if any, of the meta-ethical strategies (deontology, utilitarianism, virtue ethics etc.) leads decisions that fit with our ethical intuitions, relative to which types of agentive contexts. Similarly, by varying design and functionalities of humanoid robots philosophers can join neuroscientists in the empirically investigation which, if any, of the extant alternative accounts of our capacity of 'mind-reading' (theory of mind, simulation theory, phenomenology, mind-shaping) are most adequate and what this implies for the philosophical interpretation of mental discourse.

(1597 words)

Breazeal, C. (2002). *Designing Sociable Robots*. MIT Press.

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems*, 42:167-175.

Capurro, R., Nagenborg, M., Capurro, R. & Nagenborg, M. 2009, *Ethics and Robotics,* IOS Press, Amsterdam.

Gunkel, D. 2012. The Machine Question . MIT Press.

Coeckelbergh, M. (2012). *Growing Moral Relations: Critique of Moral Status Ascription.* Palgrave Macmillan.

Fong, T., Nourbakhsh, I., and Dautenhahn, K. 2003. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42: 143-166.

Funk, M. & Irrgang, B. 2014, *Robotics in Germany and Japan: Philosophical and Technical Perspectives,* Peter Lang International Academic Publishers.

Hakli, R. 2014. Social Robots and Social Interaction. In: Seibt et al. 2014, 105-115.

Hakli, R., Seibt, J. (eds.). 2017. *Sociality and Normativity for Robots*. Springer (forthcoming).

Larsen, S. 2016. *The As-If of Para-Social Interactions and the Robot as Fictional Character—Semantic, Ontological, and Methodological Reflections on Human-Robot Interaction Discourse*. PhD Dissertation, Aarhus University.

Nagenborg, M. 2007, "Artificial moral agents: an intercultural perspective", *International Review of Information Ethics* 7: 129-133.

Nakada, M., Capurro, R, 2013. An intercultural dialogue on roboethics. In: Nakada, M., Capurro, R. (eds.): *The Quest for Information Ethics and Roboethics in East and West*. Research Report on trends in information ethics and roboethics in Japan and the West. http://www.capurro.de/intercultural_roboethics.html, pp. 13-22.

Nørskov, M. 2011. *Prolegomena to a Philosophy of Social Robotics*. PhD Dissertation, Aarhus University.

Nørskov, M. 2015. Revisiting Ihde's Fourfold "Technological Relationships": Application and Modification. *Philosophy & Technology*, 28 (2): 189-207.

Seibt, J., Hakli, R., Nørskov, M., (eds). 2014. *Sociable Robots and the Future of Social Relations. Proceedings of Robo-philosophy 2014*, IOS Press, Amsterdam.

Seibt, J. 2014. Varieties of the 'As-If': Five Ways to Simulate an Action. In: Seibt et al. 2014, 97-105.

Seibt, J., Nørskov, M., Schack Andersen, S. (eds.). 2016. *What Social Robots Can and Should Do. Proceedings of Robophilosophy 2016/TRANSOR 2016.* IOS Press, Amsterdam.

Seibt, J. 2016. "Integrative Social Robotics—A New Method Paradigm to Solve the Description Problem *and* the Regulation Problem?" In: Seibt, J. et al. 2016, 104-119.

Seibt, J. 2017. Towards An Ontology of Simulated Social Interactions—Varieties of the 'As-If' for Robots and Humans. In: Hakli, R., Seibt, J. (eds.), *Sociality and Normativity for Robots—Philosophical Investigations*, Springer, 11-41.

Seibt, J., Hakli, R., Nørskov, M., (eds). 2017. *Robophilosophy—Philosophy of, for, and by Social Robotics*. MIT Press (forthcoming).

Sullins, J. P. (2008). Friends by design: A design philosophy for personal robotics technology. In: Peter Kroes, Pieter E. Vermaas, Andrew Light, Steven A. Moore (eds.). *Philosophy and Design*, Springer: Amsterdam, pp. 143-157.

Turkle, S. 2011. *Alone Together* . Basic Books, New York.

Vallor, S. 2011. Carebots and caregivers: Sustaining the ethical ideal of care in the twenty-first century. *Philosophy & Technology*, 24: 251-268.

Van Wynsberghe, A. 2015. *Healthcare Robots—Ethics, Design, and Implementation*. Routledge, New York.

Wallach, W., & Allen, C. 2010. *Moral machines: Teaching robots right from wrong*. Oxford University Press.