This text has been published as:

Brinck, I. 2023. Social Robots in Social Institutions: Scaling up and Cutting Back on Cognition. In: Proceedings from Robophilosophy 2022: Social Robots in Social Institutions, Helsinki, August 16-19 2022,. R. Hakli, P. Makela, J. Seibt (Eds.), pp. 615-619. Frontiers in Artificial Intelligence and Applications, Vol. 366. IOS Press BV. DOI: 10.3233/FAIA220667

Please refer to the published version!

SOCIAL ROBOTS FOR SOCIAL INSTITUTIONS: SCALING UP AND CUTTING BACK ON COGNITION

Ingar Brinck
Department of Philosophy and Cognitive Science, Lund University
Box 192 SE- 22100, Lund Sweden
ingar.brinck@fil.lu.se

Current technological change is rapid and far-reaching, more so than ever before in human history. It is transforming all dimensions of human life, leading to large-scale adaptation. Among the disruptive new technologies that are being introduced into society, social robots are distinguished by their hybrid existence between mere thing and mindful agent. They are physical machines capable of interacting with the surroundings and designed to collaborate with humans on human tasks while interacting in human ways. In contrast to rescue, delivery, and patrol robots that replace human labour, these robots take social roles such as tutor, peer, learner, companion, or assistant. On a more general note, social robots are expected to work closely with humans – as partner, colleague, family, and friend. Yet, that a robot can act *as* a social entity, does not entail the robot constitutes a social being in its own right (Fischer 2019). The question is whether social robots are capable of participating in and contributing to human social institutions such as healthcare, education, and economy; and if this is so, the follow-up question concerns what this may entail for society in the longer perspective. The emphasis of the present talk lies on the first question, the supposed contribution of social robots to social institutions. Raising a few queries concerning the ability of the BDI-paradigm and affective robotics to provide an adequate reply to this question, at least in its present formulation, I will briefly outline an alternative that lays down a new path in HRI, based in the notions of embodied, embedded, dynamic, and distributed cognition. I claim these notions are particularly well-suited for designing social institutional forms of HRI, because they permit modelling the relevant cognitive processes as unfolding in the physical space that humans and robots share. The environment provides the resources for HRI such as artefacts, routines, and embodied social norms that simultaneously constrain and enable the emergence of novel institutional practices involving human and machine.

RATIONAL AND SOCIABLE ROBOTS

Two major paradigms dominate social robotics today. The *belief-desire-intention (BDI) paradigm* (Kinny, Geogeff, & Rao 1996) is based in the formalization of folk psychology, or human everyday reasoning about behaviour, in terms of logical inferences that range over representations of the mental states of rational agents. It construes social cognition as an extension of individual cognition. Whenever possible, social concepts are derived from concepts that concern the individual (e.g., group belief is modelled on individual belief), while new concepts are added for irreducible social notions such as reciprocity, commitment, and social norm. The BDI-paradigm is used in computer simulations of human-robot interaction ranging from pairwise to multi-agent contexts and is especially valuable in large contexts, as in models of the transportation system of a big city, or traffic control in the same city during rush

hours. The simulations take place in idealized environments stripped from the unique elements that characterize interaction in the real world and are intended to generalize, even, to hold universally across situations.

The second paradigm, *affective robotics*, exploits the human penchant for empathy and social bonding to maintain engagement with the robots and has been described as a paradigm shift from intelligence to emotion. Breazeal (2003) defined the genuinely social or *sociable* robot by its communicative abilities, functionally indistinguishable from those of humans: Not used as tool or machine to perform tasks, but an agent with its own motivations and goal. Breazeal's definition set the bar for affective robotics that develops robots capable of displaying and reacting to emotion within multimodal face-face interaction (Breazeal 2003; Dautenhahn 2007). The robots regularly occur in healthcare, education, social services, and medicine, and have been shown to enhance positive feeling, strengthen self-confidence, remove stress, and relieve social anxiety (Rasouli, Gupta, Nilsen, & Dautenhahn 2022). Except for promoting task-performance, the robots are used as social mediator to reinforce children's social skill and in the treatment of ASD.

The two paradigms approach human-robot interaction (HRI) from distinct perspectives, the one rational and detached via observation, the other sociable and engaged via emotion. Specializing on different tasks, each is successful in its domain. On the other hand, it is not clear what it would mean on these accounts for robots to literally participate in human social institutions, or how such participation would occur. The regular mode of procedure in HRI is to first analyse interaction, then package it in a format that makes it controllable, which in practice means abstraction and idealisation. While this procedure may be necessary in certain contexts, it conceals important dimensions of human behaviour that justify behaviour beyond the rational and sociable.

The BDI paradigm can provide definitions of the social institutions and allows for representing the rules and regulations that organize social institutional activities and permit participation -- in principle, at least. On this view, participating in social institutions would require capacities for logical reasoning about own and others' beliefs, desires, and intentions, including about the institution in question, the ways it functions, and the behaviour and transactions it sustains and the values it embraces. Yet, such reasoning capacities would not be sufficient. Many of the tasks social robots are intended to perform in social institutions are performed together with others, not on their own. To get this right, the model will have to look further than the individual and the first-person perspective and model the processes that involve agents with each other and the environment. We need to reveal the points that agents have in common that enable complementary action, and to represent the agents in the second person that conditions acting together, as the You and I that form a temporary We, a plural, heterogenous, agent (Brinck, Reddy, & Zahavi 2017). On-line performance of a collaborative task demands real, embodied agents that are physically co-present in the human environment, capable of coordinating movements, gaze, bodily orientation, and so on, and of responding contingently to maintain involvement until the projected task or activity is achieved.

Thus, research in human-machine interaction and related areas indicates that physical embodiment is a game-changer in human-robot interaction: In comparison to embodied conversational agents, computer simulations, voice chatbots, etc., interaction with locomoting social robots yield human behaviour patterns that resemble patterns in human-human interaction to a significantly greater extent than do patterns caused by virtual agents (embodied conversational agents, agent-based simulations, voice chatbots, tele-present robots), including video replay of the same robot as in the physical condition. Humans take physically embodied robots more seriously (Fischer et al. 2019), follow their suggestions more often (Bainbridge et al. 2011), and find them more convincing (Fischer et al. 2021). Physical robot interaction results in better learning outcomes (Leyzberg et al. 2012), and a more positive perception of the robot and better user performance (Li 2015). Finally, physical co-presence is associated with diffusion of responsibility (for action), or reduced sense of agency in humans, e.g., Ciardo et al. (2020) show that ascribing intentional agency to a robot reduces the experience of causing one's own actions.[1] This reduction of sense of agency entails a diffusion of responsibility, so-called by-stander effect that indicates the human is sharing the responsibility with the robot.

[1] Diffusion of agency is related to bystander effect (e.g., Darley & Latané 1968).

Schilbach et al. (2013) report that neuroimaging and psychophysiological studies provide evidence of processing differences related to social knowing depending on whether (i) a person is a detached observer or is experiencing the situation in emotional engagement with the observed agent, (ii) the experimental paradigm allowed for interaction or not, and (iii) data collection takes place at the level of a single or of two (or more) individuals. It seems these processing differences depend on a fundamental distinction between knowing others in (embodied) engagement and knowing them by (detached) observation, i.e., at a distance, or without rapport.

Regarding affective robotics, it is not clear how the capacity for emotional interaction relates to or supports participation in social institutions– if the latter depends on the former, or rather, we are dealing with distinct forms of social interaction. Skills for emotional engagement can support the functional side of participation in social institutions, e.g., in healthcare. Yet emotional engagement is far from sufficient for coping in the institutional context.

Scenarios can be quite complex and highly structured in the multi-agent scenarios of social institutions, the structure shaping the unfolding interaction. Such scenarios are likely to be more common in social institutions than in spontaneous dyadic interaction. Trasmundi's (2012, 2019) intricate analysis of the roles, responsibilities, duties, and practices among the members of the emergency team at a Danish hospital illustrates the details of this complexity. Except for knowledge about rules, regulations, and routines, about seniority and local traditions, and about bodily skills, Trasmundi points to the importance of embodied skills such as professional vision (knowing where to look, what to attend to and when, and how to attend, and moreover, the capacity to grasp the significance of what you perceive), habitual movement patterns in the confined space of the ward shared by the team members, sensitivity to the others' bodily expressions of emotion including ignorance and wariness, to local and institutionally coded norms of proximity, and much more. Cognition is *both* embodied and embedded in the higher-order social relationships that exert top-down influence on face-face interaction.

Clearly, contributing to social institutions requires abilities for interaction equally on the microscale of milliseconds, the macroscale of decades, and the mesoscale that lies in-between. It presupposes the sharing of habits, tradition, and conventions, and a wide variety of contextual knowledge and skills that employees develop incrementally throughout their working-life. Colleagues will have to act as a team in the first-person plural of the We and temporarily disregard their status as individuals. The least hesitance as to whether you act as one of the team and representing the institution, or individually, pursuing your own private goals, can be detrimental to the outcome of any institutional processes and activities – in the practice, decision-making and policy development of medicine, education, healthcare, and so on.

While most robots operate in controlled environments where the task space does not change, social robots are intended to operate in real environments, which are dynamic and continuously change, and consequently difficult to restrict. To compensate for the loss of control that ensues from having the real world as task space, social robots sometimes are tele-operated (remotely controlled). The operation of semi-supervised social robots is restricted before they are taken in use, which means such robots can operate without on-line human control. Constraints can be built-in, or pre-programmed, or based in pre-training on customized datasets. Because the constraints are pre-designed and depend on old data, they do not enable novel behaviour. This means semi-supervised robots cannot respond to new types of input or new situations: Responding would have required updating the constraints.

A major cause behind the long-standing problems of humanoids is the incapacity to adapt to the real world that is dynamic and continually changing. The behaviour of robots trained on stationary data samples does not apply in novel circumstances and will cause the robots to stall. Unable to extrapolate learning, such robots tolerate only minor variations among sensor input.

Incremental learning and on-line adaptation to variations in the real world is not possible with traditional machine learning. Increasing the scope for mismatch by using deep machine learning constitutes an improvement but does not solve the problem. It is costly in terms of processing and therefore fails on account of sustainability and might be considered a last resort. However, using non-linear algorithms that enable continuous learning and personalisation we can design unsupervised, autonomous robots that are capable of learning from experience, and of developing new behaviour in response to environmental variation without previous pre-training, improving the quality in time.

INTERACTION IN THE CULTURAL CONTEXT

Arguably, designing HRI for the societal and cultural contexts will benefit from modelling cognition as fundamentally social, relational, and multi-scalar, allowing for processes on different scales to inform and support each other. Moreover, to avoid the problem of programming *all* the knowledge required for coping in the real world by a companion or teaching robot, and the ensuing problem of applying general knowledge to the specificities of the present, we might discard the traditional way of modelling knowledge and cognition as rational, observational, and first-person.

Suppose instead we conceive of social cognition as *distributed* (Hutchins 1995; Kirsh 2013). This would permit robot(s) and human(s) to share the task space in the real world and use it to support learning and joint action. To illustrate, a particular configuration of the task space would permit memorizing the various elements of the task and how they are interrelated, and furthermore the order in which you need to access them to realize the task. Additionally, aspects of the task space can function as cues to action that help you to find the way, or as landmarks that signal in what direction you will find the goal and then can provide additional, local information that benefit wayfinding (Prasad et al. 2020). Modelling cognition as distributed in shared space will increase the transparency of social interaction and facilitate participation. Furthermore, it will make embodied cognitive processes directly actionable for the other participants. This tends to increase trust among the members of a group, because it means that cognitive processes can be interrogated and modulated publicly, e.g., halted, countered, or developed. Finally, the distributed cognition model will reduce processing and increase sustainability.

I suggest letting the robot

(i) develop the knowledge appropriate for task performance in situ, while learning to perform the task in real life and in collaboration with humans, say, with the person who needs a robot companion and the nurse that will oversee the robot interaction, or at school, with the teacher and the parents of the pupils that will benefit from robot tutoring,

(ii) learn together with the humans while mutually adjusting, and tailor learning differently to different humans,

(iii) learn incrementally and continuously like humans do, refining and tweaking learning during performance.

These abilities are likely to make social robots better equipped to participate in social institutions than the ones the BDI-paradigm and affective robotics entail, linking the development of practical knowledge and institutional praxis to the material, social, and cultural context of the task. For instance, personalisation (see point (ii)) can be expected to increase user experience and make the robot seem less intrusive while it still can maintain an institutional role or function (Churamani et al. 2017, 2020). Moreover, the distributed model will permit users to participate in the robot's learning from the outset, instead of facing a finished product as is the case with social robots today, the design of which to a great part is based in models of individual cognition. However, humans are heavily influenced by those they interact with and the contexts of interaction in which they participate. The distributed model would involve social learning with robot and human functioning as resources for each other, and engage them in continuous learning, creating the conditions for continual improvement of their relation, much like humans do who plan to work together over longer periods (e.g., Lesort et al. 2020; Peternel et al. 2014).

REFERENCES

Bainbridge, W.A., Hart, J.W., Kim, E.S. et al. 2011. The benefits of interactions with physically present robots over video-displayed agents. International Journal of Social Robotics, 3, 41–52.

Breazeal, C. 2002. Designing sociable robots. Cambridge Mass: MIT Press.

Brinck, I., Reddy, V., Zahavi, D. 2017. The primacy of the We. Cambridge Mass: MIT Press, pp. 131-148.

Churamani, N., et al. 2017. The impact of personalisation on human-robot interaction in learning scenarios. Proceedings of the 5th International Conference on Human Agent Interaction HAI '17, Bielefeld, Germany, pp. 171–180, New York, NY, USA: ACM.

Churamani, N., Barros, P., Gunes, H., Wermter, S. 2020. Affect-driven modelling of robot personality for collaborative human-robot interactions. arXiv:2010.07221.

Ciardo, F., Beyer, F., De Tommaso, D., Wykowska, A. 2020. Attribution of intentional agency towards robots reduces one's own sense of agency. Cognition 194:104109.

Dautenhahn, K. 2007. Socially intelligent robots: dimensions of human–robot interaction. Philosophical Transactions of the Royal Society B: Biological Sciences 362(1480), 679–704.

Darley, J. M., Latané, B. 1968. Bystander intervention in emergencies: Diffusion of responsibility. Journal of Personality and Social Psychology 8, 377–383.

Fischer, K. 2019. Why collaborative robots must be social (and even emotional) actors. Techné: Research in Philosophy and Technology 23(3), 270-289.

Fischer, K., Jensen, L. C., Zitzmann, N. 2021. In the same boat. The influence of sharing the situational context on a speaker's (a robot's) persuasiveness. Interaction Studies 22(3), 488-515.

Hutchins, E. 1995. Cognition in the wild. Cambridge, MA: The MIT Press.

Kinny, D., Georgeff, M., Rao, A. 1996. A methodology and modelling technique for systems of BDI agents. In: Van de Velde, W., Perram, J.W. (Eds.) Agents breaking away. MAAMAW 1996. Lecture Notes in Computer Science, vol. 1038. Springer, Berlin, Heidelberg.

Kirsh, D. 2013. Embodied cognition and the magical future of interaction design. ACM Transactions in Computer-Human Interaction 20, 1.

Lesort, T. et al. 2020. Continual learning for robotics: definition, framework, learning strategies, opportunities, and challenges. Information Fusion 58, 52-68.

Leyzberg, D., Spaulding, S., Toneva, M., Scassellati, B. 2012. The physical presence of a robot tutor increases cognitive learning gains. Proceedings of the annual meeting of the cognitive science society, 34.

Li, J. 2015. The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. International Journal of Human-Computer Studies 77, 23–37.

Peternel, L., Petrič, T., Oztop, E., Babič, J. 2014. Teaching robots to cooperate with humans in dynamic manipulation tasks based on multi-modal human-in-the-loop approach. Autonomous Robots 36(1–2), 123–136.

Prasad, A., Sharma, B., Vanualailai, J., Kumar, S. 2020. Stabilizing controllers for landmark navigation of planar robots in an obstacle-ridden workspace, Journal of Advanced Transportation, Article ID 8865608, 1-13.

Rasouli, S., Gupta, G., Nilsen, E., Dautenhahn, K. 2022. Potential applications of social robots in robot-assisted interventions for social anxiety. International Journal of Social Robotics 14(5),1-32.

Schilbach, L., Timmermans, B., Reddy, V., et al. 2013. Toward a second-person neuroscience. Behavioral and Brain Science 36, 393–414.

Sandini, G., Sciutti, A. 2018. Humane robots—from robots with a humanoid body to robots with an anthropomorphic mind. ACM Transactions in Computer-Human Interaction 7, 1.

Sciutti, A., Mara, M., Tagliasco, V., Sandini, G. 2018. Humanizing human-robot interaction: on the importance of mutual understanding. IEEE Technology and Society Magazine 37, 22–29.

Trasmundi, S. B. 2012. Interactivity in health care: bodies, values and dynamics. Language Sciences 34(5), 532–542.

Trasmundi, S. B. 2019. Skilled embodiment in emergency medicine: the "interactivity turn" and its implication for theory and practice. Chinese Semiotic Studies 15(4), 627-651.