What does it take to be a social agent?

Víctor Fernández Castro, Raul Hakli, Aurélie Clodic

What does it take to be a social agent?

Víctor FERNÁNDEZ CASTRO[a,b1][OBJ], Raul HAKLI[c] and Aurélie CLODIC[a]

[a] *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France*
[b] *Institut Jean Nicod, CNRS UMR 8129, DEC, ENS, PSL University, Paris, France*
[c] *University of Helsinki, Helsinki, Finland*

Abstract. The aim of this paper is to present a philosophically inspired list of minimal requirements for social agency that may serve as a guideline for social robotics. Such a list does not aim at detailing the cognitive processes behind sociality but at providing an implementation-free characterization of the capacities and skills associated with sociality. We employ the notion of intentional stance as a methodological ground to study intentional agency and extend it into a social stance that takes into account social features of behavior. We discuss the basic requirements of sociality and different ways to understand them, and suggest some potential benefits of understanding them in an instrumentalist way in the context of social robotics.

Keywords. Social agency, intentional stance, social stance, normativity, motivation

1. Introduction

The rapid advance of social robotics along with its pervasive use in all contexts of public and private social interaction raise a number of fundamental questions about the possible impact of social robots on human social practices and the possibly irreversible interventions in the spaces of human culture and sociality [1]. These questions are especially relevant in a context where social robots are often equipped with capacities for expressing social signals or for exhibiting behavioral tendencies that boost concern, feeling and care on humans. Such robotic behavior may automatically create in humans a reaction to reciprocate or to treat a robot as a fully autonomous social being. Reactions like that may be partially explained by the holistic nature of our perception of sociality. We find it difficult to interpret a particular behavior as being social without implicitly attributing a set of capacities at the same time. So, we tend to "socialize" the robot in an unrealistic manner, that is, we attribute to it a number of capacities that typically go together in the social beings we typically encounter [2,3]. For instance, when a smartphone or a robot greets us in a happy sounding voice, it may evoke in us an emotional reaction or an attempt to make an eye-contact which are typical reactions in human social interaction. If the phone or robot cannot respond in a reciprocal and socially appropriate manner, our expectations are not met and we may feel disappointed or frustrated. Added to this problem is the fact that we probably do not have a complete model of the list of capabilities that we associate with this tendency to socialize, and therefore cannot reliably assess the consequences of the use of social robots for our culture and values. Identifying a set of features that are necessary for sociality might alleviate the possible negative impacts, for example, by enabling robots with alternative

---

[1] Corresponding Author, Víctor Fernández Castro. LAAS-CNRS, 7 Avenue du Colonel Roche, 31400 Toulouse; France e-mail: vfernandezcastro@gmail.com.

behavior that is socially effective but minimally invasive, thereby avoiding exploiting human emotions and feelings and harming human integrity or autonomy.

The aim of this paper is to take a first step in trying to propose a possible solution to this problem. To this end we will present a philosophically inspired list of minimal requirements for social agency. This list is intended to be a guide to the type of behavioral properties or capabilities instrumentally characterized without appealing to implementation specifications that every social agent, we believe, should meet.

In what follows, we will employ a method initiated by Daniel Dennett who suggested that human beings have a natural tendency to understand each other as intentional agents. According to Dennett, when we try to understand other people's behavior, we adopt an intentional stance toward them, meaning that we understand them as agents that have intentional states, basically beliefs and desires, and that are rational in the sense that they select actions that according to their beliefs, best serve the satisfaction of their desires [4]. This is a useful approach for understanding folk psychology and explains also our tendency to attribute intentional states to animals and artifacts like robots. It also explains the holistic nature of intentional attitudes: It does not make sense to attribute only one type of intentional attitude to a system, like only beliefs or only desires, because they only function together in the agent's processes of practical reasoning.

However, the intentional stance does not yet include the features that we are interested in this paper. It only has to do with intentional states and rational agency, but it does not have anything to say about the features specific to social agency. As argued e.g. by Michael Tomasello et al., human social and cultural cognition requires more than understanding intentional actions and attitudes [5]. Nevertheless, we do not want to go so far as to adopt a "personal stance" toward robots that would treat robots as persons with such attributes as consciousness, linguistic abilities, and fitness to be held morally responsible for their actions [6,7], nor a "cultural stance" that would require understanding of social institutions and well-developed capacities for collective intentionality and cooperation [5]. Hence, we are looking for a set of capacities that would be required of minimally social agents, capacities that presuppose intentional agency but would not yet suffice for full personhood or understanding of culture.

Our work might be seen as preliminary work for establishing such a concept of social agents that would include human beings, social animals, and social robots. This concept would establish a set of features that roboticists may use as a starting point in designing social robots before going to richer sets of sociality features [8]. It would also be useful for users because informing them about such sociality features that still fall short of personhood would help them to avoid making unrealistic assumptions about robots' capacities in terms of consciousness, emotions, or morality. Adopting an instrumentalist understanding of the stance approach, moreover, allows us to legitimately talk about "social robots" without having to commit ourselves to the idea that robots "really" are social in the sense humans are, or share with us similar mechanisms that produce the types of behavior we consider social. In the next section, we will discuss the stance approach in more detail. Then we will consider the various types of behaviors and capacities required of social agency.

2. From the intentional stance to the social stance

Consider social robots and the question what it would mean for a robot to be social? What in general is it that makes something social? Let us try to answer this question

along the lines of Dennett's reply to the question of what makes something intentional. Dennett famously proposed three stances that one can take when encountering a novel object and trying to understand and predict its behavior: The physical stance, the design stance, and the intentional stance. In the physical stance, we try to understand an object based on its physical properties. If I examine a dry, old branch fallen from a tree, I can predict, simply on the basis of the material properties of the piece of wood, that it will break if I apply force to it and burn if I set fire to it.

However, if inside a building I encounter a lift, I don't know enough of its material properties to be able to predict what will happen if I go inside. I will have to take a design stance: A lift is a device built for a specific purpose, namely for transporting people and stuff between storeys in a building. Equipped with this knowledge, I can infer that most probably it will take me to another storey. If it doesn't start moving immediately, I may begin to wonder how I can indicate to which storey I'm planning to go, and again use the design stance to infer that the numbered buttons on its wall are probably there for that purpose. Of course, if I were an electrical engineer specialized in lift control technology, I might be capable of understanding the functioning of a lift completely on the basis of its physical parts. Still, it would be much more cumbersome to couch an explanation using merely physical descriptions of the system, without referring to its purpose, or to the purposes of its parts.

When encountering animals or other people, neither the physical stance nor the design stance works well, because I don't have enough information about their material composition nor their purpose. However, with some background knowledge and perhaps some observation of their behavior I may understand them as agents, as intentional beings who have beliefs and desires about the world. Adopting such an intentional stance, I can make sense of their behavior: This rabbit is running away from me because he thinks I am a threat and he wants to avoid threats. This person is entering this building because she wants to get home and she believes she lives in this building. Attributing such intentional states to animals and human beings makes sense, because this way of thinking makes their behavior understandable and easier to predict than taking a physical or a design stance. Again, it might not be impossible to couch an explanation or prediction using one of these other stances, but it would quite certainly be too complicated to be practically useful.

Sometimes explaining and predicting actions from the intentional stance becomes difficult or cumbersome too. Why is this guy hurrying to get to that office building so early in the morning? Does he really want to do that, he looks so tired? And why is he stopping at the red light if he wants to get there as soon as possible? Does he really want to stand there all of a sudden? In fact, why is everyone stopping at red lights? Do they enjoy watching red color or what? When we observe how human beings behave, it starts to seem that they are not just trying to satisfy their desires in the light of their beliefs. Or at least, trying to explain their actions in terms of their desires and their means-end beliefs soon becomes quite complicated. At this point, it might be useful to extend our explanatory vocabulary once again, and start using vocabulary referring to social norms, social roles, and social institutions. We can say that the guy is hurrying because he has a job in that building, and that in this role he is expected to be there at 8 am, and that is a norm in their company. But he is stopping at the red light, just like everyone else, because it is a rule in the society to stop at red lights. At this point, we are employing what we call a social stance.

A social stance may become useful when the explanation of an agent's behavior doesn't seem to conform to instrumental belief-desire patterns in a straightforward way. Consider a robot, proceeding in a straight path across the room towards the exit, suddenly changing its path like it were avoiding a collision with two people discussing in the

middle of the room although there would be plenty of room to go between them. Why did it not choose the shortest path? A possible hypothesis would be that the robot was aware of a social norm stating that one should not walk in between people who are having a conversation because it would disturb them. If such reference to social aspects consistently improves our ability to explain and predict an agent's behavior, our social stance will be justified and we can consider the agent as a social agent. Hence, if reference to such social features proved to be essential to explaining and predicting a robot's behavior, it would be legitimate to call it a "social robot".

How seriously should we take the attribute "social" here? Are we saying that robots can "really" be social agents or just that it is a useful way of talking about them? Many have viewed Dennett's intentional stance approach as an instrumentalist move instead of a serious account of intentional agents with beliefs and desires, but Dennett himself has denied such interpretations [4]. However, we will here treat the stance idea in an instrumentalist fashion that allows that there might be much more to both intentionality and sociality than what interpretation of other agents suggests. We avoid committing to a behavioristic endeavor of reducing psychological explanation to a list of behavioral patterns. We are not negating that behind the manifestation of skills and actions that helps us to identify a social agent as such there are different complex and sophisticated psychological devices. Our aim is merely trying to understand and explain talk about "social" robots: Just like we have a tendency to see goal-directed behavior everywhere and attribute intentionality to even small moving figures in the computer screen [9], we tend to attribute social capacities based on certain types of behavior. In this sense, the approach aims to make sense of the talk of social robots and legitimize it as a useful way of speaking. However, it admits that this kind of sociality may differ from human sociality in that it does not presuppose that the agents have similar abilities or similarly functioning mechanisms. In that sense, it allows sociality between agents with asymmetric capacities although it doesn't distinguish between various ways and degrees in which the agents' capacities may differ [2,10]. Moreover, it can provide social robotics with a sufficiently abstract characterization of sociality that allows different implementation variants, avoiding that the design process is biased or constrained by psychological or cognitive science theories and their details about how these skills and abilities are realized or implemented in specific cognitive structures.

What then are the types of behavior for an object to display or the capacities that an agent should have in order for the social stance to be appropriate? Let us next consider some candidates.

3. The requirements for sociality

Imagine that we are observing some object and try to understand its behavior. We try to explain and predict what it does from both the physical stance and the design stance. However, our explanations turn out very complicated and our predictions go wrong. Hence, we adopt an intentional stance and try to attribute to it intentional states like beliefs and desires and see if we could see it as a rational agent trying to pursue its goals. If this makes it easier to predict and explain its actions, it is reasonable to take the object to be an intentional agent. This is the first step: In order to count as a social agent, the object must be an agent in the first place. This motivates the intentional stance requirement below:

> *Intentional stance requirement:* In order to be considered as an agent, the target of observation must behave in a way that its behavior can be interpretable as being rationally motivated by beliefs, desires, and intentions.

As we said above, one may wonder whether an intentional agency requires real mental states that, according to some theories, involve phenomenological experiences that robots may not experience. However, the intentional stance requirement is neutral with respect to any theory about the nature of mental states or even any theory of intentionality as long as being an intentional creature involves behaving in such a way that the relevant course of action can be understood by postulating beliefs, desires and intentions to the actor[2].

Social stance assumes that the agent is not merely intentional but also social, implying, among other things, that the agent is capable of differentiating between other agents and mere objects, with a tendency to alter its behavior in the presence of other agents. This characterization comes close to Max Weber's account of social action [11], according to which "Action is 'social', insofar as its subjective meaning takes account of the behavior of others and is thereby oriented in its course". In this sense, we could say that, in a minimal sense, the social agency requires being sensitive to other agents:

> *Sensitivity to other agents:* In order to be social, an agent must have the capacity to identify other agents and be sensitive to their presence by varying one's behavior accordingly.

Sensitivity to other agents is necessary for social interaction, because social interaction requires being able to predict the actions of others and to coordinate one's own actions with theirs. It is a minimal requirement that can involve very different types of cognitive capacities at different levels. For instance, one agent exhibits sensitivity to other agents when they are able to unconsciously mimic a target behavior [12] or to involuntarily synchronize their behavior thanks to subconscious motor coordination [13]; but also, when they engage in high-level cognitive reasoning directed to the mental states of other agents, so-called mindreading or folk psychology. A system with such a mindreading capacity can be called a second-order intentional system in Dennett's terminology [4], that is, a system able to understand other agents' behavior as being rationally motivated by beliefs and desires, and to be capable of applying the intentional stance to themselves and have beliefs about their own beliefs and desires. Social animals like chimpanzees have been reported to be capable of understanding others as intentional agents and to be able to use this understanding to manipulate others' experiences and actions in order to achieve their own goals [14].

Proper sociality requires more than a capacity to manipulate others. It requires a tendency to seek contact with others, to interact with them, and to form relationships with them. Hence, we must introduce another requirement for social agency:

---

2 At this point, someone could argue that the requirement is not neutral with respect to our material knowledge of the system. In other words, the fact that we know that robots are artifacts with a different functioning from that of human beings could interfere with our ability to understand them as intentional agents. In response, one could argue that we normally assume that non-human animals are intentional beings of some kind even when we know that their psychological functioning is different from ours. In addition, as we have pointed out above, we seem to attribute intentionality to even small moving figures in the computer screen [8]. Finally, the objection lies on an empirical claim that could surely be dependent on the evolution of technology in robotics: Even if we discover that human beings are reluctant to attribute intentionality to robots on the basis of their material composition, nothing assures that this could not change with the technological development of robots.

> *Tendency for interaction requirement:* social agents must be able to exhibit a tendency to engage in interaction with other agents and manifest certain interest in others' actions.

In humans and other social animals, this exhibition of tendencies to be interested in others and interact with them probably depends on a conglomerate of diverse motivations. Several authors argue that many types of prosocial behavior depend on individualist and selfish inclinations, so they are oriented, for instance, to protect individual reputation [15]. However, it can be argued [16,17] that, although the reputation management mechanism is a strong source of motivation, it can hardly be a general explanation for prosocial tendencies. Moreover, it is doubtful that reputation management could be developmentally prior to general dispositions to engage in positively valenced interactions with other people within a framework of long-lasting concern for each other's welfare [16]. Whatever the mental structures behind human social motivation are, what we can be sure of is that an object that can be called social in a minimal sense must, at least, be capable of exhibiting certain behavioral profiles that we can associate with an inclination or tendency to seek social encounters and care about the behavior of others and, at least in some cases, find social interactions intrinsically rewarding.

From unconscious synchronization to intentional coordination as well as communication and cooperation, we can find a very diverse array of capacities that reflect different ways to interact with other agents. For instance, communication is an integral part of most social interaction: In order to engage in social interaction, one must understand others as intentional agents and to somehow signal one's own intentions to them, and it has been argued that this requires a capacity for third-order intentionality [6]. However, several authors [18,19] have argued that the context where social coordination and understanding depend on attributing mental states to each other is quite rare and that we often outsource our capacity for prediction and coordination in social and cultural norms, rules and scripts (for a similar point in robotics see [20,21]). According to the so-called mindshaping or regulative view, then, an important part of our high-level sensitivity to other agents lies in generating expectations regarding the social norms and values that populate our social niche.

This view leads us to another central requirement for social agency, that is, normative understanding. Even though there arguably can be sociality without normativity [22,23], the role of norms in human and probably non-human [24] sociality is so pervasive that it is doubtful that we could consider an agent as social without attributing certain sensitivity to norms to her. Social agents recognize that there are social norms and rules that apply to them. Merely behaving as social and cultural norms dictate does not suffice, but social agents seem to also have certain sensitivity or understanding of how other agents' behavior relates to such normative structures and expectations. In this sense, we think that at least some minimal kind of normativity is a necessary requirement for sociality. We introduce the requirement as follows:

> *Normative understanding requirement:* social agents must be capable of understanding their and other agents' behavior under the light of the expectations generated by social norms, rules and conventions, and modifying their own behavior to accord with them.

Normative understanding is a complex capacity that has generated a substantial amount of literature in philosophy [25]. However, as in the previous requirements, one may find distinct levels of complexity regarding normative understanding, since complex architecture involving different mechanisms for understanding behavior derivate from

norms or motivating normative actions [26] to so-called naïve normativity [24], attributable to non-human animals and based on basic capacities for identification of agency, in-group/out-group sensitivity and responsiveness to appropriateness. In other words, this requirement merely demands a social agent to be able to recognize actions or courses of behavior as conforming to a norm (i.e. that the action fulfills the expectations generated by a norm), and to regulate their behavior accordingly.

Social norms and rules have a strong conventional component, which implies that it is doubtful that they could have emerged in every individual independently. This suggests an important connection between normativity and sociality in general but also between normativity and the pro-active capacity of social agents to evaluate each other and mutually influence and regulate each other's behavior. In addition, the basic function of norms and rules seem to be the improvement of coordination so that it is unlikely that they could be spread among a population without the ability of individuals to assess and influence the behavior of other actors in light of the expectations generated by the rules and norms. In this sense, a social agent must be able to evaluate the behavior of others and to react to norm violations. Thus, when someone violates our normative expectations (either moral, social, or rational), we exercise different reactions to acknowledge the failure: we blame, sanction, ask for reasons, and so on. The mindshaping or regulative view mentioned above argues that our capacity of evaluating and reacting to counternormative behavior explains the homogenization of human behavior in a population in accordance with a set of norms and values. That means that normative understanding is tied to having a pro- or con-attitudes toward the action or the agent and that such evaluation often triggers different regulative behavior toward the target action. As a result, social agency is also connected to a requirement for evaluation and regulation:

> *Normative evaluation and regulation requirements:* social agents must be able to evaluate positively and negatively a particular course of action in accordance with a norm and trigger different appropriate regulative behavior like sanctioning responses, asking for reasons, blaming, manifesting approval, or justifying the action

Normative regulation presupposes a capacity to learn and motivate normative standards, a tendency to form normative expectations, and to engage in regulative behavior when others' behavior deviates from the standards, such as teaching and sanctioning others [27–29]. We have exemplified such regulation with sophisticated responses like blaming or asking for reasons. However, regulation can also be exercised through low-level punishing actions such as emotional responses like anger or sadness or merely avoiding cooperation or reciprocation. Be that as it may, social agents' capacities for evaluating and regulating each other are manifestations of the so-called second-personal relation in which they acknowledge the other as an other, as an agent that is in some sense similar to oneself and who has certain value and demands respect [7,30].

The notion of evaluation and regulation seems to imply certain affective or motivational aspects. After all, we would hardly understand why a creature may expend energy and time in evaluating and regulating other actions without presupposing certain motivation to do so. Human social behavior does not only reflect cognitive sophistication but important and complex motivations to engage in social interactions. Humans spend a significant amount of their time engaged in social situations carrying out cooperative projects and interacting with each other and generally prefer company to solitude. These motivations are manifested in different aspects, for instance, humans exhibit emotions like embarrassment, guilt or shame, whose nature is profoundly social and they

"necessarily depend on other people's thoughts, feelings or actions, as experienced, recalled, anticipated or imagined at first hand, or instantiated in more generalized consideration of social norms or conventions" [31: 131]. In other words, such emotional profiles are premised on the fact that humans profoundly concern themselves with others' thoughts and actions. Moreover, humans' decisions can be strongly influenced by others' expectations about their behavior [15] and they are demonstrably cooperative even when they do not have reasons to expect further interactions where they can receive compensation [32].

Something like the above requirements seem to us to provide sufficient and jointly necessary conditions for social agency in a minimal sense. We believe there is a class of social agents that are capable of rudimentary social interaction that still falls of short of full social interaction that assumes symmetrical capacities from participants and this is what we have tried to characterize. There are further requirements that have been suggested as requirements for personhood, moral agency, or sophisticated social and cultural agency exhibited by human beings, that we have not included. For instance, Dennett mentions consciousness and capacity for verbal communication as conditions of personhood [6]. Consciousness grounds many features that are essential for human-like communication, such as the capacity to experience emotions, but at least according to philosophical thought experiments concerning zombies, consciousness is not a property that manifests itself in behavior and hence we have excluded it from the list of behavioral features that warrant taking a social stance. As to verbal communication, in our view it does not seem to be necessary for social behavior as many animals can interact socially with more rudimentary communicative capacities than what could be considered verbal. The same applies to sophisticated capacities for collective intentionality that ground human-like cooperation and joint action: As Tomasello et al. have argued [5], these features distinguish human beings from other great apes, but since we are here interested in sociality more broadly, we have excluded them from our list.

4. Conclusion

The aim of this paper was to present a list of requirements for sociality that could inspire social robotics in its attempt to design more robust and socially efficient robotic agents. To do that, we have considered an instrumentalist approach that attempts to highlight the basic capacities for sociality while remaining neutral to the details of implementation behind such capacities. This instrumentalism may provide designers more liberty to play with the physical features of the robot without being limited or restricted to the widespread idea that robotic sociality must mirror human sociality. Moreover, we have speculated with the idea that the instrumentalist approach may help to create robots that exhibit social traits without necessarily triggering emotional engagement on the part of the human being, and thus, less invasive.

Acknowledgment

References

[1] Seibt J, Damholdt MF, Vestergaard C. Integrative social robotics, value-driven design, and transdisciplinarity. Interaction Studies. 2020;21(1):111–144. Available from: https://doi.org/10.1075/is.18061.sei.

[2] Seibt J. Towards an Ontology of Simulated Social Interaction: Varieties of the "As If" for Robots and Humans. In: Sociality and Normativity for Robots. Springer International Publishing; 2017. p. 11–39. Available from: https://doi.org/10.1007/978-3-319-53133-52.

[3] Young JE, Sung J, Voida A, Sharlin E, Igarashi T, Christensen HI, et al. Evaluating Human-Robot Interaction. International Journal of Social Robotics. 2010 Oct;3(1):53–67. Available from: https://doi.org/10.1007/s12369-010-0081-8.

[4] Dennett D. Intentional Systems Theory. The Oxford Handbook of Philosophy of Mind. Oxford University Press; 2009. Available from: https://doi.org/10.1093/oxfordhb/9780199262618.003.0020.

[5] Tomasello M, Carpenter M, Call J, Behne T, Moll H. Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences. 2005 Oct;28(5):675–691. Available from: https://doi.org/10.1017/s0140525x05000129.

[6] Dennett D. Conditions of Personhood. In: What Is a Person? Humana Press; 1988. p. 145–167. Available from: https://doi.org/10.1007/978-1-4612-3950-5_7.

[7] Laitinen A. Robots and Human Sociality: Normative Expectations, the Need for Recognition, and the Social Bases of Self-Esteem. Frontiers in Artificial Intelligence and Applications. 2016;290 (What Social Robots Can and Should Do): 313–322. Available from: https://doi.org/10.3233/978-1-61499-708-5-313.

[8] Baraka K, Alves-Oliveira P, Ribeiro T. An Extended Framework for Characterizing Social Robots. In: Springer Series on Bio- and Neurosys-tems. Springer International Publishing; 2020. p. 21–64. Available from: https://doi.org/10.1007/978-3-030-42307-02

[9] Heider F, Simmel M. An Experimental Study of Apparent Behavior. The American Journal of Psychology. 1944;57(2):243–259. Available from: http://www.jstor.org/stable/1416950

[10] Seibt J. Classifying Forms and Modes of Co-Working in the Ontology of Asymmetric Social Interactions (OASIS). Frontiers in Artificial Intelligence and Applications. 2018;311 (Envisioning Robots in Society - Power, Politics, and Public Space): 133–146. Available from: https://doi.org/10.3233/978-1-61499-931-7-133.

[11] Weber M. Economy and Society. English translation, original German edition 1921. Berkeley, CA: University of California Press; 1978

[12] Chartrand TL, Bargh JA. The chameleon effect: The perception–behavior link and social interaction. Journal of Personality and Social Psychology. 1999;76(6):893–910. Available from: https://doi.org/10.1037/0022-3514.76.6.893.

[13] Richardson DC, Dale R, Kirkham NZ. The Art of Conversation Is Coordination. Psychological Science. 2007 May;18(5):407–413. Available from: https://doi.org/10.1111/j.1467-9280.2007.01914.x.

[14] Tomasello M. A natural history of human thinking. Cambridge, Massachusetts: Harvard University Press; 2014.

[15] Heintz C, Celse J, Giardini F, Max S. Facing expectations: Those that we prefer to fulfil and those that we disregard. Judgment and Decision Making. 2015 September;10(5):442–455. Available from: https://ideas.repec.org/a/jdm/journl/v10y2015i5p442-455.html.

[16] Fernández Castro VF, Pacherie E. Joint actions, commitments and the need to belong. Synthese. 2020 Jan; Available from: https://doi.org/10.1007/s11229-020-02535-0.

[17] Godman M, Nagatsu M, Salmela M. The Social Motivation Hypothesis for Prosocial Behavior. Philosophy of the Social Sciences. 2014 May;44(5):563–587. Available from: https://doi.org/10.1177/0048393114530841.

[18] Andrews K. Understanding Norms Without a Theory of Mind. Inquiry. 2009 Oct;52(5):433–448. Available from: https://doi.org/10.1080/00201740903302584.

[19] Zawidzki TW. The function of folk psychology: mind reading or mind shaping? Philosophical Explorations. 2008 Sep;11(3):193–210. Available from: https://doi.org/10.1080/13869790802239235.

[20] Fernández Castro V. Shaping Robotic Minds. Frontiers in Artificial Intelligence and Applications. 2014;273 (Sociable Robots and the Future of Social Relations):71–78. Available from: https://doi.org/10.3233/978-1-61499-480-0-71.

[21] Fernández Castro V. Mindshaping and Robotics. In: Sociality and Normativity for Robots. Springer International Publishing; 2017. p. 115–135. Available from: https://doi.org/10.1007/978-3-319-53133-5_6.

[22] Brandl JL, Esken F. The Problem of Understanding Social Norms and What it Would Take for Robots to Solve it. In: Sociality and Normativity for Robots. Springer International Publishing; 2017. p. 201–215. Available from: https://doi.org/10.1007/978-3-319-53133-5_10.

[23] Satne G. What Binds Us Together. Philosophical Topics. 2014;42(1):43–61. Available from: https://doi.org/10.5840/philtopics20144213.

[24] Andrews K. Naïve Normativity: The Social Foundation of Moral Cognition. Journal of the American Philosophical Association. 2020;6(1):36–56. Available from: https://doi.org/10.1017/apa.2019.30.

[25] Smith B C. Meaning and rule-following. In: Routledge Encyclopedia of Philosophy. Routledge; 2016. Available from: https://doi.org/10.4324/9780415249126-u021-1.

[26] Sripada CS, Stich S. A Framework for the Psychology of Norms. In: The Innate Mind: Volume 2: Culture and Cognition. Oxford University Press; 2007. p. 280–301. Available from: https://doi.org/10.1093/acprof:oso/9780195310139.003.0017.

[27] Fernández Castro V. Regulation, Normativity and Folk Psychology. Topoi. 2017 Nov;39(1):57–67. Available from: https://doi.org/10.1007/s11245-017-9511-7.

[28] McGeer V. Mind-making practices: the social infrastructure of self-know in gagency and responsibility. Philosophical Explorations. 2015 May;18(2):259–281. Available from: https://doi.org/10.1080/13869795.2015.1032331.

[29] Zawidzki T. Mindshaping : a new framework for understanding human social cognition. Cambridge, MA: MIT Press; 2013.

[30] Brinck I, Balkenius C. Mutual Recognition in Human-Robot Interaction: a Deflationary Account. Philosophy & Technology. 2018 Dec;33(1):53–70. Available from: https://doi.org/10.1007/s13347-018-0339-x.

[31] Hareli S, Parkinson B. What's Social About Social Emotions? Journal for the Theory of Social Behaviour. 2008 Jun;38(2):131–156. Available from: https://doi.org/10.1111/j.1468-5914.2008.00363.x.

[32] Chaudhuri A. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. Experimental Economics. 2010 Sep;14(1):47–83. Available from: https://doi.org/10.1007/s10683-010-9257-1.