In press at *AI and Society*

You and I, robot

Shaun Gallagher
Lillian and Morrie Moss Chair of Excellence in Philosophy
University of Memphis (USA)
Research Professor of Philosophy and Cognitive Science
School of Humanities
University of Hertfordshire (UK)

Abstract: I address a number of issues related to building an autonomous social robot. I review different approaches to social cognition and ask how these different approaches may inform the design of social robots. I argue that regardless of which theoretical approach to social cognition one favors, instantiating that approach in a workable robot will involve designing that robot on enactive principles.

Keywords: social robots, social cognition, theory of mind, theory theory, simulation theory, interaction theory, enactive cognition.

Introduction

Researchers in advanced robotics are attempting to build autonomous social robots that will be able to seamlessly and reliably interact with humans in specific situations.[1] This project motivates both philosophical and practical questions about what precisely is required in a robot if it is to be able to engage in something close to the kind of interaction that characterizes human-human relations, even if only on a pragmatic level. That is, if we set aside concerns that have to do with the complicating factors of care and emotion and focus simply on communicative capacity, is there some guiding ideal (in the sense of a Kantian ideal that we could aim at, even if we are not convinced that we can achieve it) for designing and building such a robot?

In setting aside questions about care and emotion I don't mean to suggest that these issues are not important for human or human-robot interactions, or that they are not solvable. Even now it seems possible to build robots that elicit care and certain emotions from humans (e.g., Kismet, see Breazeal 2002). Even if the robot is not designed, in terms of its appearance and behavior (e.g., facial features, vocal intonations, etc.), to elicit care from a human responder, it seems possible that humans may come to care for a robot

[1] I'm involved in a large project of this sort, although my contribution is in the distant theoretical background and is focused on questions about the nature of gesture and the possibility of building gesture into the repertoire of a robot's communicative skills. My research on robotics is supported by a grant from the Robotics Collaborative Technology Alliance and General Dynamics, #64018180, Social cues and behaviors in HR collaboration. Also, thanks to the Marie Curie Initial Training Network, Towards an Embodied Science of Intersubjectivity (TESIS). Marie Curie Actions, European Commission Research for support of my research on intersubjectivity.

in a significant way to the extent that they come to depend on it. This sometimes happens in regard to other machines – automobiles, computers, etc. It's an open question, however, whether this kind of caring for a machine is the same kind of caring, and simply a matter of a difference of degree from that which humans experience for each other. It's also an open question whether caring can go the other way, that is, whether a robot can have anything more than a pragmatic care for a human (in the sense of simply taking care of that human). Also, setting these questions aside doesn't mean that care and emotion do not enter into and shape our everyday human interactions in important ways. Stripping away care and emotion from our everyday interactions may in fact change them in essential ways, since we are not simply pragmatic agents. Still, for many purposes one does not have to enter into caring or emotional relations with others who may simply be playing pre-defined social roles in highly specified contexts (e.g., the person who sell me coffee at the airport).

In more sustained contexts, where humans interact with robots over a significant amount of time and in a diversity of situations, what should the nature of that interaction be? One central issue in this circumstance is whether we (the robot and I) can understand one another – something that may be accomplished by specific communicative practices. Communicating with a robot via speech and/or gesture, however, turns out to be a complicated thing if one aims at smooth and reliable communicative practices. Much of communication depends on implicit aspects – what is *not* said is sometimes of greater importance than what is explicitly said, and non-conscious gestures, postures, movements and bodily expressions are often more important than consciously produced signs. For this and other reasons there is a vast and growing literature on social cognition in psychology, neuroscience, and philosophy of mind. In the following sections I'll review different approaches to social cognition and ask how these different approaches may inform the design of social robots.

Social cognition

Under the title of 'theory of mind' (ToM), two models of social cognition have dominated the debate in this literature: Theory theory (TT) and simulation theory (ST). According to TT, our observations of another's behavior together with our folk-psychological theory, are the basis for making inferences about the other's mental states. Since we have no direct access to the other person's mental states, understanding others must be a form of 'mindreading' or 'mentalizing' based on an ability to use folk psychology to infer the other's beliefs and desires.

ST, in contrast, suggests that we do not have to appeal to folk psychological theory because we can use our own mind as a model to simulate the beliefs and desires of others. Alvin Goldman (2005), for example, describes the simulation process in three steps.

> First, the attributor creates in herself pretend states intended to match those of the target. In other words, the attributor attempts to put herself in the target's 'mental shoes'. The second step is to feed these initial pretend states [e.g., beliefs] into some mechanism of the attributor's own psychology … and allow

> that mechanism to operate on the pretend states so as to generate one or more new states [e.g., decisions]. Third, the attributor assigns the output state to the target …" [e.g., we infer or project the decision to the other's mind]. (Goldman 2005b, 80-81.)

Importantly, the simulator is working off-line. The beliefs and desires generated in the simulation are pretend beliefs and desires, which are then projected onto the mind of the other person. This form of mindreading may be the result of a conscious process (so-called high-level simulation) or the outcome of sub-personal mechanisms (e.g., neurons and so-called low-level simulation) (Goldman 2006).

There are various problems with each of these approaches, hence the ongoing debate between them. One problem that pertains to both approaches, however, is a particular version of the frame problem. I refer to it as the 'starting problem' (Gallagher 2011a&b), and one can see it in the very first step of the simulation routine as Goldman describes it. According to ST the first step is for me to create "pretend states intended to match those of the target." But how do I know what states will match those of the target? That is supposedly the problem that ST is attempting to solve. It seems that in order to run a simulation which would allow me to know the states that would match those of the other person, I already have to know what those states are. In regard to TT, in order to generate the correct inference about the other person's mental states, I would have to appeal to the appropriate piece of theory or rule of folk psychology; but to do that I would already have to understand something about the other person's situation and their mental states.[2]

A recently developed alternative to these ToM approaches is interaction theory (IT). IT appeals to evidence from developmental psychology, phenomenology and embodied and enactive approaches to cognition. On this view, our primary way of understanding others depends not on inference or simulation routines, but on second-person embodied interactions in practical and social contexts. As we interact with others we can perceive their meanings and their intentions, as well as their emotions, in their bodily movements, gestures, facial expressions, in what they are looking at and what they are doing in the rich pragmatic and social contexts of everyday life. IT conceives of intentions, for example, not as mental states to which I have no direct access, but as something that is built into the structure and style of your actions, which I can see. In most everyday circumstances, I can see what you want to do in the way you are doing it, and in many cases, this is obvious to me because I am not simply observing your behavior in an off-line (third-person) mode, but am engaged with you in an on-line (second-person) interaction. Moreover, in our interactions I understand the meanings of your actions (and gestures and expressions, etc.), enactively, that is, in terms of social affordances, in terms of my possible responses to your actions. In many cases, also, the particular situation (the physical setting, the social environment) we are in does some of the work. The meaning of a certain gesture or a certain action is specified by the situation in which it is

[2] ST based on low-level processes associated with mirror neurons (see, e.g., Gallese and Sinigaglia 2011 for a recent statement) has its own set of problems that I will not try to rehearse here (see Gallagher 2007).

enacted; the meaning of the same set of movements may be specified differently in a different situation. Accordingly, in most of our everyday circumstances, mindreading, and the attempt to explain other people's behaviors in terms of mental states that are hidden away in their minds, are ordinarily not required. The information that we have from our embodied interactions and the surrounding context is sufficient for understanding (see Gallagher 2005; 2008 for further discussion of IT).

Importantly, instead of looking for specific mechanisms (e.g., mirror neurons or ToM mechanisms) in an individual in order to explain his or her ability to mindread, IT claims that in some circumstances interaction itself (the enactive engagement of two or more individuals which is not reducible to the actions of the individuals qua individuals) constitutes social cognition (De Jaegher, Di Paulo and Gallagher 2010). Interaction in this regard is something like the tango – a meaningful event that emerges as something more than simply the *addition* of movements made by two individuals moving as individuals. In this regard there is good evidence that interaction involves dynamic action patterns with precise (although variable) timing. For example, in a set of contingency experiments, Murray and Trevarthen (1985) have shown the importance of live interaction between mother and two-month old infant. They use a double TV monitor experiment where mother and infant interact by means of a two-way live television link. The infants engage in lively interaction in this situation. When the infant is presented with the recorded replay of their mother's actions, however, the infant quickly disengages and becomes distracted and upset. This change in behaviour occurs even though the visual stimulation is exactly the same, although now lacking contingency with the infant's movements. These results have been replicated, eliminating alternative explanations such as infants' fatigue or memory problems (Nadel *et al.* 1999; Stormark and Braarud 2004).

Evidence in developmental studies also indicates that there is no starting problem involved in the interaction of social cognition. In most cases, we are not off-line, attempting to find a way to start understanding the other; rather, from very early in infancy, we are drawn into second-person social interaction in the embodied practices of primary intersubjectivity (Trevarthen 1979). We grow up in these practices and find ourselves already in them in ways that allow us to recognize meaning, intention, and emotion in the other's embodied behavior. Even in instances when we are off-line observers, our embodied experiences with others, which give us easy access to what has been called the 'massive hermeneutical background' provided by those early and continuing experiences with others, provides the starting point for our understanding.[3] In effect, if there is a role in our everyday encounters for mindreading of the TT or ST sort, what solves the starting problem for such practices is this background provided I part by our ongoing embodied intersubjective interactions.

[3] The phrase is from Bruner and Kalmer (1998). See Gallagher (2011b). There is more to the story than I can discuss here. The massive hermeneutical background is provided not only by our bodily interactive skills, but also by the wealth of narratives that we have at our disposal (see Gallagher and Hutto 2007; Hutto 2008).

Social robotics

How should the ongoing debates in the field of social cognition, across the lines that divide TT, ST, and IT, inform the design of social robots? Clearly, a TT robot is going to require a different design than an ST robot. A ToM robot of either sort, however, would seemingly depend on off-line processes designed to regard humans in social contexts as problems to be solved in terms of a logic of folk (belief-desire) psychology. Mindreading robots would have to compute the mental states (or what philosophers call 'propositional attitudes') of others, using observed behaviors as inputs. One might think that building a TT or ST robot would be easier than building an IT robot, since running computations based on off-line observation and a logic of mental state inference or matching state simulation would seemingly be simpler than building a robot capable of an on-line understanding of facial expressions, social context, etc., and thereby capable of interacting with humans. I'll argue that this is not so because anything like the TT or ST robot will already have to be an IT robot if it is required to function in the seamless and relatively reliable way that characterizes most human interactions.

It's not clear, for example, how such TT or ST robots would solve the starting problem if they did not already have the massive hermeneutical background that humans gain through their embodied interactions. This background includes (and at certain points in development, *involves having*) not just knowledge of the factual kind, but skills and practical know-how, a good sense of sensory-motor contingencies, perceptual capacities that have been attuned by years of practice to nuances in facial expressions, postures, styles of movement, etc. (including those that tell us about the mood or emotion of the other person), and the ability to understand and draw on narratives (Gallagher and Hutto 2006; Hutto 2008). Having and being able to draw on this background depends not just on individual ability, but requires being in and having long exposure to situations that are already situations of social interaction. That is, if social cognition were simply a matter of an individual (human or robot) having the right internal mechanism rather than already being engaged in embodied interaction with others, and enacting processes that depend on being with others for their proper development, then there would still be a question of how in any particular circumstance the individual would know what rule to follow, or what simulation to make.

A ToM robot is not necessarily impossible, but the starting problem would likely slow it down and make it unreliable. Some evidence of this comes from the reports of people with high-functioning Autism or Asperger's Syndrome. It's not that they lack ToM, as frequently claimed (e.g., Baron-Cohen 1995); it's that ToM is the only thing that they do have in this regard. They report using procedures well described by TT – mind reading by making inferences based on observed behaviors (Zahavi and Parnas 2003). As Oliver Sacks indicates, describing the well known case of Temple Grandin,

> This implicit knowledge, which every normal person accumulates and generates throughout life on the basis of experience and encounters with others, Temple seems to be largely devoid of. Lacking it, she has instead

> to "compute" others' intentions and states of mind, to try to make algorithmic, explicit, what for the rest of us is second nature. (Sacks 1995, 258; also see Blackburn et al. 2000).

At best, then, a ToM robot would be something like an autistic robot. In such robots, the timing and timeliness of response would be off, and not only would this mean poor performance on their part, but it would also interfere with our ability to understand and interact with them, since attunement to the other person in interaction is a dynamic process that involves precise responsive timing on both sides.

In other words, even if our goal is to build TT or ST robots, it seems that we would first have to build them as IT robots. Both folk-psychological theoretical inference and any form of the simulation process will depend on starting with behavior that is recognizable and understandable in an embodied (primary-intersubjective) way. In addition, if the robot were not capable of entering into the precise dynamics of interaction, in a way that is understandable to a human, it would be difficult for the human to start up the inferential or simulational processes.

One might argue, from a TT perspective, that one doesn't need to appeal to *folk* psychology in such cases, but to a specialized *robot*-psychology. That is, one would only need to know the limited number of rules that the robot would be expected to follow in its behavior. A robot with a limited repertoire of behavior, however, would not be a sophisticated autonomous social robot of the sort that we are trying to characterize. It would resemble a tool robot rather than an autonomous robot that could operate in varied and complex situations. Indeed, if the TT robot operated only on a set of algorithms small enough to allow us to predict its limited behavior, then it's not clear that it would have sufficient resources to mindread us. A TT robot capable of understanding human mental states would have to have sophisticated mindreading skills, and that would imply that its own propositional attitudes – and therefore robot psychology – would be more complex.

Likewise, one problem for an ST robot is that it would have to be sufficiently like us to be able to simulate our mental states. ST depends on this resemblance or ability of the simulator to match the target. For the ST robot to resemble us sufficiently to support its ability to simulate behavior, however, it would also need to be sophisticated enough to engage in that behavior. Yet it would not be sufficient to design a robot that simply and automatically matches or imitates our motor or mental states, as suggested by the mirror neuron version of ST. The kinds of responses required for everyday interaction do not reduce to, and are more subtle and complex than simple replication or imitation. That is, in our everyday engagements with others, we are not simply mirroring their actions; we are enactively responding with actions that may be complementary, or helpful, or oppositional, or, more generally, responsive.

Accordingly, both TT and ST lead to the idea that a robot capable of smooth and reliable interaction with humans would already have to be an IT robot, that is, a robot that behaves sufficiently like us so as to engage in embodied interaction. On the one hand, if

this were not the case, TT and ST would not be able to solve the starting problem. On the other hand, if this is the case, that is, if the robot does behave sufficiently like us so as to engage with us in embodied interaction, then not only is the starting problem solved, but, according to IT, the problem of social cognition in most instances is solved without necessarily employing theoretical inference or simulation.

These considerations suggest that the guiding ideal for building a sophisticated autonomous social robot should be that we build such a robot according to the principles of IT, rather than TT or ST. This is the case even if our preferred theory is TT or ST, or a hybrid theory that maintains that social cognition may involve different kinds of performance in different kinds of situations. A sophisticated autonomous social robot will have to be a robot built on enactive principles for embodied interaction. Is this possible? Not only do we not want to rule anything out *apriori*, but, in fact, research in evolutionary robotics provides minimal instances of such interactive machines.

Using techniques of evolutionary robotics Di Paolo (2000) evolved agents that were able to achieve coordination through interaction. Two robots, whose only task was to locate each other and remain close as they moved through a large space, used simple auditory signals and rotating motor behavior, to set up a specific sound pattern that differentiated between self and non-self, and simplified what would otherwise be a complex recognition problem. By these means the agents were able to accomplish their task. Importantly, when an individual agent was presented with a recording of its partner from the previous successful interaction, it was unable to reproduce its own behavior due to the non-contingency involved in the recording. One-sided coordination was not achievable, which suggests the important contribution of interaction itself, a finding similar to the Murray and Trevarthen (1985) contingency experiments mentioned above.

Similar results were found in another minimal behavior, perceptual crossing, first tried with humans and replicated in artificial agents (Auvray, Lenay and Stewart 2009; Di Paolo, Rohde and Iizuka 2008). Blindfolded, using a computer mouse, two human agents find each other's icons on a wrap-around line, distinguishing the other agent from similar fixed objects and shadow (inactive) agents, through a spontaneously emerging strategy (a particular back-and-forth movement) that works only through the interaction of both agents. Again using evolutionary robotics Di Paolo et al. (2008) show that the differentiation between fixed object and the other agent is essentially dependent on timing. Scanning (moving back and forth on) the fixed object involves a longer duration than scanning the other agent since the other agent is also scanning. This translates into a difference of perceived size (the fixed object appears larger or longer than the other moving agent). As Di Paolo, Rohde, and De Jaegher (2011) point out, the smaller perceived size of the other agent depends on the two agents moving in an antiphase pattern which they must coordinate in interaction. A failure in the precise timing in the interaction (within some set boundaries) would result in a failure of task.

Such experiments with minimal behaviors in evolutionary robotics can help to identify the enactive principles involved in embodied interaction. The challenge is to see if these principles scale up to the kind of non-minimal behaviors that will characterize full-out

human-robotic interactions in worldly environments. If so, it then remains an open question whether one would need to include additional layers of TT or ST architecture, or whether an IT robot would be sufficient for most instances of second-person, you-and-I relations.

References

Auvray, M., Lenay, C. & Stewart, J. (2009). Perceptual interactions in a minimalist virtual environment. *New Ideas in Psychology*, *27*(1), 32-47.

Baron-Cohen, S. 1995. *Mindblindness: An essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press.

Blackburn, J., K. Gottschewski, E. George, and L. Niki. 2000. A discussion about theory of mind: From an autistic perspective. *Proceedings of Autism Europe's 6th International Congress*, Glasgow, 19–21 May 2000. Available from URL: http:/ /www.autistics.org/library/AE2000-ToM.html.

Breazeal, C. 2002. *Designing Sociable Robots*. Cambridge, MA: MIT Press.

Bruner, J. and Kalmar, D. A. 1998. Narrative and metanarrative in the construction of self. In M. Ferrari and R. J. Sternberg (eds.), *Self-Awareness: Its Nature and Development* (pp. 308-331). New York: Guilford Press.

De Jaegher, H., Di Paolo, E. and Gallagher, S. 2010. Does social interaction constitute social cognition? *Trends in Cognitive Sciences,* 14(10), 441-447.

Di Paolo, E. A. 2000. Behavioral coordination, structural congruence and environment in acoustically coupled agents. *Adaptive Behavior* 8: 27-47.

Di Paolo, E. A., Rohde, M. and De Jaegher. 2010. Horizons for the enactive mind. In J. Stewart, O. Gapenne, and E. A. Di Paolo (eds.), *Enaction: Toward a New Paradigm for Cognitive Science* (33-87). Cambridge, MA: MIT Press.

Di Paolo, E. A., Rohde, M. & Iizuka, H. 2008. Sensitivity to social contingency or stability of interaction? Modelling the dynamics of perceptual crossing. *New Ideas in Psychology*, *26*(2), 278-294.

Gallagher, S. 2011a. In defense of phenomenological approaches to social cognition: Interacting with the critics. *Review of Philosophy and Psychology.* DOI 10.1007/s13164-011-0080-1

Gallagher, S. 2011b. Narrative competency and the massive hermeneutical background. In Paul Fairfield (ed.), *Hermeneutics in Education* (21-38). New York: Continuum

Gallagher, S. 2008. Inference or interaction: Social cognition without precursors. *Philosophical Explorations,* 11 (3), 163-73.

Gallagher, S. 2007. Simulation trouble. *Social Neuroscience,* 2 (3-4), 353-65.

Gallagher, S. 2005. *How the Body Shapes the Mind*. Oxford: Oxford University Press.

Gallagher, S. and Hutto, D. 2008. Understanding others through primary interaction and narrative practice. In: J. Zlatev, T. Racine, C. Sinha and E. Itkonen (eds.), *The Shared Mind: Perspectives on Intersubjectivity* (pp. 17-38). Amsterdam: John Benjamins.

Gallese, V. and Sinigaglia, C. 2011. What's so special about embodied simulation? *Trends in Cognitive Sciences* 15 (11), 512-519.

Goldman, A. I. 2005. Imitation, mind reading, and simulation. In: S. Hurley and N. Chater (eds.), *Perspectives on imitation II* (pp. 80-91). Cambridge, MA: MIT Press.

Goldman, A. I. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. New York: Oxford University Press.

Hutto D. D. 2008: *Folk Psychological Narratives: The Socio-Cultural Basis of Understanding Reasons*. Cambridge, MA: MIT Press.

Murray, L. & Trevarthen, C. (1985). Emotional regulations of interactions between two-month-olds and their mothers. In T. M. Field & N. A. Fox (Eds.), *Social perception in infants* (pp. 177-197). Norwood, NJ: Ablex Publishing.

Nadel, J., Carchon, I., Kervella, C., Marcelli, D. & Réserbat-Plantey, D. 1999. Expectancies for social contingency in 2-month-olds. *Developmental Science*, *2*(2), 164-173.

Sacks, O. 1995. *An Anthropologist on Mars*. New York: Vintage Books.

Stormark, K. M. & Braarud, H. C. 2004. Infants' sensitivity to social contingency: a double video study of face-to-face communication between 2- and 4-month-olds and their mothers. *Infant Behavior & Development*, *27*(2), 195-203.

Trevarthen, C. B. 1979. Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. Bullowa (ed.), *Before Speech* (pp. 321-48). Cambridge: Cambridge University Press.

Zahavi, D., and J. Parnas. 2003. Conceptual problems in infantile autism research: Why cognitive science needs phenomenology. *Journal of Consciousness Studies* 10, no. 9–10:53–71.