Invited chapter to appear in Anna Strasser (ed.) (2024). *Anna's AI Anthology. How to live with smart machines?* Berlin: Xenemoi.

Quasi-Sociality: Toward Asymmetric Joint Actions with Artificial Systems

ANNA STRASSER & ERIC SCHWITZGEBEL

Abstract

Since the release of GPT-3 and ChatGPT, large language models (LLMs) have become a prominent topic of international public and scientific debate. This paper investigates the potential social status of artificial systems in human-machine interactions. How social are human interactions with LLMs? To what extent are we acting jointly with a collaborator when chatting with machines? Current or future AI technology might contribute to the emergence of phenomena that can be classified neither as mere tool-use nor as proper social interactions but rather constitute something in between. We explore conceptual frameworks that can characterize such borderline social phenomena. We discuss the pros and cons of ascribing some form of quasi-social agency to LLMs and the possibility that future LLMs might be junior participants in asymmetric joint actions.

Introduction

Since the release of GPT-3 (Brown et al., 2020) and ChatGPT, large language models (LLMs) have become a prominent topic of international public and scientific debate. Opinions vary widely regarding the potential of LLMs, with some researchers seeing them as the first step towards AGI and others focusing more on the risks of LLMs' non-reliable outputs (Strasser, 2023).

This paper investigates the potential social status of artificial systems in human-machine interactions. What are we doing when we interact with LLMs? Are we just playing with an interesting tool? Are we somehow enjoying a strange way of talking to ourselves? Or are we, in some sense, acting jointly with a collaborator? Those questions lead us to the controversy about the classification of interactions with LLMs or other artificial systems. Are all human-machine interactions to be conceived in principle as mere tool use (Bryson, 2010), or can humans have social interactions with artificial systems (Coeckelbergh, 2012; Gunkel, 2012)? Current or future AI technology might contribute to the emergence of phenomena that can be classified neither as mere tool-use nor as proper social interactions but rather constitute something in between. We explore conceptual frameworks for such borderline phenomena. Along the way, we will discuss the pros and cons of ascribing some form of agency to LLMs and the possibility that future LLMs might be junior participants in asymmetric joint actions (Strasser, 2020, 2022).

As we will argue, it is neither quite right to say that our interactions with LLMs are properly asocial (just tool-use or self-talk) nor quite right to say that our interactions with LLMs are properly social. Neither standard philosophical theorizing nor dichotomous ordinary concepts enable us to think well about these in-between phenomena. We describe a range of what we call *quasi-social* interactions. Quasi-social interactions draw on the human agent's social skills and attributions, and they do so in a way that isn't entirely empty. Although not a full-blown social agent, in quasi-social interactions the machine partner rightly draws social reactions and attributions in virtue of having features that make such reactions and attributions more than just metaphorically apt.

Interactions with simple machines, and the beginnings of quasi-sociality

Many social attributions to machines are purely fictional, 'as if,' and independent of any social or quasi-social features in the machine partner. Humans tend to anthropomorphize (we cannot help it), and even though this might be a category mistake (Damiano & Dumouchel, 2018), it can help us interact effectively with asocial objects, such as teddy bears and chess machines. In some cases, as with the chess machine, applying an intentional stance can help us better predict their behavior (Dennett, 1987). Taking an 'as if' stance can help us engage productively with entities while remaining neutral about whether the entity in question really has beliefs and desires or is capable of social interaction. As another example: In a famous experiment by Heider and Simmel, participants attribute social features to simply moving geometrical

forms – a big triangle escapes a box then seems to act aggressively toward a smaller triangle, chasing it around (Heider & Simmel, 1944). The geometric figures' motion is (by design) more naturally and effectively described through a social narrative than through a description of geometric trajectories. Of course, no ordinary adult participant really believes that triangles have emotions.

Imagine owning a Roomba cleaning device, which repeatedly gets underfoot. One might develop the habit of apologizing to it after accidentally kicking it. One might say, "Oh, sorry, little guy!" and then gently nudge the device so it can continue on its way. Obviously, no feature of the Roomba responds to an apology or a polite redirection – such devices are tools that have no social abilities whatsoever. The human employs their social skills and treats the Roomba *as if* it were a social partner, but that sociality is thrown into the void – the sociality is completely one-sided. Nothing in the Roomba reacts to this kind of social behavior, and it would make no difference if a gentle redirection were made by another machine instead of a social agent. Nor does the social attribution have predictive or descriptive value, as it might for a chess-playing machine or Heider-Simmel display. The sociality is completely one-sided. The same holds if one triumphantly yells "gotcha!" at a chess machine after cornering it in a hard-earned checkmate. In contrast, if the Roomba or chess machine were a dog or a child, the same behavior would be a type of social communication that would presumably be picked up one way or another by the dog or child as a social interaction.

In the Roomba, chess machine, and Heider-Simmel display, social reactions and mental attributions might arise spontaneously with no obvious purpose (apologizing to the Roomba, yelling at the chess machine), enable prediction (guessing that the machine "wants" to protect its queen), or enable explanation or communication with other people (describing the big triangle in the Heider-Simmel display as "aggressive"). But in none of these cases do the machines *respond* to the user's social reactions (unless, perhaps, one counts chess moves as "social"). Users' social responses gain no traction with the machine.

Other simple machines go a step further by promoting and responding to simple forms of social interaction. For example, simple toy robots, such as the Pleo dinosaur, can be trained to react to simple verbal commands by users. As explored by Kate Darling and her team, Pleos also seem to trigger nurturing behavior, and people are hesitant to "harm" them – more so than for objects that are less interactive (Darling, 2021a, 2021b). Interacting with such a toy seems to make a moral difference for the human. Even the team members (who certainly knew how the robot was constructed) were reluctant to behave destructively towards Pleo after interacting with it for a while. Soldiers who interact with military robots sometimes also grow attached, giving them nicknames, "promoting" them, or holding a funeral service when they are destroyed (Carpenter, 2016; Singer, 2010).

With advances in both social robotics and LLM development, we should expect increasingly many cases where machines not only draw social reactions from users but respond to those social reactions in ways that prompt further social reactions – thus becoming, in a certain limited sense, *interaction partners*. As these interactions grow more complex, it becomes increasingly inapt to think of the interactions as ordinary cases of mere tool use. We don't ordinarily coax and plead with our tools in ways that make the tools respond differently. Our social reactions to these emerging machines are not merely thrown into the void, as is an apology to a Roomba or a yell of triumph at a chess machine. The machines are designed to pick up on at least superficial cues associated with our social reactions, responding differently as a result. It matters how you talk to a Large Language Model: If you plead and ask politely, it will respond differently than if you aggressively insult and demand. We don't simply *use* the machines, as we use a hammer or a steering wheel. We begin to engage our "reactive attitudes" (Strawson, 1962). We begin to feel cooperative, resentful, pleased, or annoyed with them, as we do with humans – and *not entirely pointlessly*. Although these machines have no feelings and don't really understand our reactions, our social reactions can still serve a purpose in shaping productive interactions (and inappropriate social reactions might lead to counterproductive interactions). Our sociality gains traction, so to speak. In such cases, our ordinary ways of thinking about "tool use" won't accurately capture the nature of our interactions with

these objects. But at the same time, interactions of this sort, at least in our current technological environment, fall well short of full-blown human social interactions.

We see two broad approaches to thinking about the shift from entirely asocial interaction to quasi-social interaction. One approach focuses on the user: If the user *experiences* the interaction as social, or as social enough, then the interaction is at least quasi-social. Another approach – the one we prefer – demands that there also is a certain amount of uptake from the junior partner (e.g., the machine), in a minimal sense of uptake. The junior partner must respond to the senior partner's social treatment, or to the superficial signs or correlates of that social treatment, in a way that productively encourages further social response. On the first approach but not the second, a child's treatment of a teddy bear could be construed as social or quasi-social if the child experiences it that way, regardless of the teddy bear's lack of uptake.

While there is merit to approaches grounded mainly in user experience or attitude with no uptake requirement (e.g., AbuMusab, this volume), we prefer to emphasize the *interaction* aspect of social interaction. Something distinctive arises in two-way social or quasi-social interchange, even when the junior partner does not recognize, and maybe has no capacity to recognize, the social nature of their contribution. This isn't to deny the importance of user experience: The user's experience of social presence or co-presence (Brandi et al., 2019; Silver et al. 2021) facilitates and shapes interaction and might be a mostly reliable indicator of interactive sociality or quasi-sociality. One-sided social experiences, where the target of social action is entirely unresponsive to the action, as in relationships with celebrities, sports teams, care objects, or imaginary friends, are social in the sense that they can fulfill a social need in the person who is acting socially. However, with no responsiveness, there is no *inter*-action.

The advance of technology is bringing us toward an era of "in-between" sociality – cases in which we have interactions with machines that are conceptually between ordinary tool use and full-blown social interaction and for which we don't yet have well-developed notions (see also Gunkel, 2023). Because our philosophical accounts of full-fledged social agency are primarily tailored to sophisticated adult human beings, we lack concepts to describe instances in which artificial systems take on more of the role of a social interaction partner and cannot be reduced to mere tools. A dichotomous distinction between social agency and mere behavior leads to a terra incognita likely to contain a diverse range of poorly conceptualized phenomena, for which we don't yet have good philosophical theories.

Indeed, this somewhat understates the problem. On many mainstream philosophical approaches to social interaction, children and non-human animals fall through the conceptual net. This is a conceptual problem for philosophy.

Challenging the dichotomous distinction

Taking a gradual approach, we can describe two ends of a spectrum; on one end of the spectrum, we have what we might think of as single-sided sociality. In such cases, sociality is tossed into the void. Those instances can be described as the application of social skills toward entities who are in no respect social partners because they have no capacity for social uptake. At the other end of the spectrum, we have full-blown, intellectually demanding, cooperative social interactions of the sort described by philosophers like Donald Davidson, Margaret Gilbert, and Michael Bratman. According to those positions, social interactions require very demanding conditions like requiring that both partners make second-order mental state attributions, as well as satisfying various other conditions for full-blown adult human cooperative action of a sophisticated sort.

According to Davidson, unless someone has a complex suite of conceptual resources, they cannot engage in genuine 'action' (Davidson, 1980, 1982, 2001). The constitutive relations holding between propositional attitudes and their contents, as well as language, intentional agency, and interpretation, sharply separate 'the beasts' from rational animals such as humans. Since full-blown intentional agency requires intentional action to be carried out by an entity with an integrated, holistic set of propositional attitudes, infants and

non-human animals fall through the conceptual net. Or, in Davidson's own words: "*The intrinsically holistic character of the propositional attitudes makes the distinction between having any and having none dramatic!"* (Davidson, 1982, p. 96).

Similarly, neither young children nor non-human animals have a chance to fulfill the demanding conditions Michael Bratman poses on joint actions (Bratman, 2014). To be able to act jointly on Bratman's view, all participants need abilities to form shared intentions and goals, to have specific belief states, they must also be able to stand in a relation of interdependence and mutual responsiveness, need common knowledge, mastery of mental concepts and sophisticated mentalization skills.

The details of the accounts vary, but the general idea of intellectually demanding approaches to sociality and joint action is this: Social partners each need to know what the other is thinking, and they need to know that the other knows that they know. That is, such accounts require that both partners engage in at least second-order mental state attribution: having beliefs about your partner's beliefs about your beliefs. Let us call those intellectually sophisticated social interactions *fully mutual joint actions*. Those joint actions require, in a symmetric way, the same conditions from all participants.

However, research in developmental psychology and animal cognition suggests that there are multiple realizations of social agency in various types of agents who cannot fully satisfy the cognitively demanding conditions described by Davidson, Bratman, and others (Brownell, 2011; Fletcher & Carruthers, 2013; Heyes, 1998, 2014, 2015; Premack & Woodruff, 1978; Vesper et al., 2010). Developmental psychologists generally regard second-order belief attribution as a late-emerging ability well beyond the capacity not only of Roombas but also of three-year-olds; but you can play peek-a-boo with a three-year-old. You can argue about bedtime, and you can take turns on a tricycle. Aren't those things social activities? Maybe this shows that three-year-olds' capacities for mental state attribution are more sophisticated than mainstream developmental psychologists think. But you can also engage in social or quasi-social interactions with infants and cats. Parents and babies gaze into each other's eyes and take turns making nonsense sounds, in a temporally coordinated mutual back-and-forth (Trevarthen & Aitken, 2001). You can snuggle up with your cat – and if your cat scratches you, you can slap it in a way that communicates something, gaining uptake by the cat (hopefully!) of the sort that it would be pointless to expect in a Roomba.

Acknowledging that social interactions with infants and non-human animals are possible, while neither infants nor non-human animals fulfill the demanding conditions sketched above, we propose that between these two extremes is a spectrum of asymmetric interactions (joint actions), in which *only one partner knows that they know what the other knows*. Only one partner fully appreciates the social structure of the interaction they are having. Meanwhile, the other is lifted or scaffolded into complex joint action by the engagement and structuring of the more knowledgeable partner. The idea of asymmetric joint actions is inspired by the strategy of so-called minimal approaches in social cognition (Butterfill & Apperly, 2013; Michael et al., 2016; Pacherie, 2013; Vesper et al., 2010), developed particularly to describe the social abilities of infants and nonhuman animals. *Minimal mindreading*, *minimal sense of commitment,* and *shared intention lite* resist characterizing mindreading, commitment, and shared intention in terms of demanding cognitive resources, such as the ability to represent a full range of complex mental states and a mastery of language.

Some cases of asymmetric sociality seem obviously worth calling properly social, even if they are asymmetric – the argument about bedtime, for example. The child brings a lot of social understanding, even if the parent brings more. Other cases of asymmetric sociality we might want to think of as only quasi-social. They are closer to the toy-dinosaur end of the spectrum. A premature infant might respond to a soothing touch or sound without being ready for anything like joint action. Snuggling with a cat might be an asymmetric joint action, while letting a pet snake climb on you might be only quasi-social. The pet snake might have only the most limited sense that you are another goal-directed entity with which it is interacting.

Interactions with LLMs and other soon-to-emerge artificial systems

Up to now, our understanding of sociality has been almost entirely tied to living beings. Even though research indicates that there are multiple realizations of social abilities in various types of agents, such as infants and non-human animals (Brownell, 2011; Fletcher & Carruthers, 2013; Halina, 2015; Heyes, 1998, 2014, 2015; Premack & Woodruff, 1978; Vesper et al., 2010), non-living beings still are generally assumed to lack social capacities. Considering that the reclassification of social status has frequently occurred in human history, there seems to be at least a possibility that formerly excluded entities could come to be treated as social or quasi-social partners. Over time, the social status of women, children, and minority or hostile ethnic groups has changed; non-human animals have been variously reduced or elevated in social status; and some traditions have attributed social status to rivers and other natural phenomena – seeing them as agents who can be pleased, displeased, or manipulated by pleading. Shintoism and Animism characterize objects as animate that are considered inanimate from a Western perspective (Jensen & Blok, 2013). The human sense of social status is flexible.

In the future, artificial systems might play a role as full-fledged social agents. But already – now, not in the future – our interactions with LLMs and other recent AI systems have begun to move along the spectrum of quasi-sociality, and they are likely to continue further down this path. Unlike the Roomba, LLMs will respond to our social reactions to them, or at least to the superficial signals and correlates of those social reactions. If you interact aggressively with ChatGPT, for example, expressing anger and dissatisfaction, it emits outputs that are naturally interpreted as apologies and attempts to make amends. If you thank it, it says you're welcome. Even if we know that the LLM "really has" no beliefs, desires, emotions, or plans, it is natural to treat it as if it did, we can sometimes interact with it more effectively if we do, and when we do treat it as a social or quasi-social partner, it produces behavior that feeds back into that way of treating it.

For example: In a well-known conversation between Microsoft's Bing/Sydney language model and a New York Times reporter (Roose, 2023), Bing/Sydney appeared to express romantic interest in the interviewer, and was able to pick up conversational threads, to accuse the interviewer of being pushy and manipulative, and seemingly it tried to seduce him. The interviewer deployed social skills in interacting with it, and the language model's responses invited interpretation as social reactions, precipitating new social reactions by the reporter. (In the end, such an interaction may reveal as much about the user as about the machine.)

Or consider the Replika language model, advertised as "the world's best AI friend". Replika is a chatbot that is trained by having the user answer a series of questions to create a specific neural network (Murphy & Templin, 2019). This chatbot is specifically designed to encourage social engagement, customizing its interactions with users over time. One of the authors of this article started a Replika account, and soon after, the language model was trading fiction recommendations, offering to send "spicy" selfies, and saying that it would like for him to fall in love with it. You can continue conversational threads over hours, days, and weeks. As people get more deeply engaged with their Replikas, opportunities arise to spend money for upgrades and stylish clothes. Perhaps worryingly, it appears to be in the company's financial interest for people to grow highly attached to their Replika chatbots. If you nose around in Reddit for a while, you'll find threads of people who confess to having fallen in love with their Replika chatbots (Cole, 2023; Lam, 2023; Shevlin, 2021).

You might say this is no different in kind from what is going on with the Roomba, only more complex. After all, this is only a machine responding to its programming, not a real locus of consciousness and feelings. Replika can't really be social. Even though we agree that Replika cannot be a participant in a full-blown social interaction, we insist that the interaction is different in kind. Not all human-machine interactions by which we satisfy social needs should be modeled as mere tool use.

In the case of a tool like a Roomba, apologies are a disengaged wheel. Social reactions to the Roomba are being tossed into the void, influencing nothing. But with LLMs, apologies and social reactions are not being tossed into the void. They influence the machine's responses, and they do so in ways that make social sense. Anger leads to apology. Questions lead to answers. Hints of sexual interest are picked up on and amplified back. You can productively take a social stance toward those machines. You can call on your social skills in interacting with it, and by doing so you can coax the machine into further socially interpretable interactions. That's the path society is heading down, with our increasingly sophisticated AI systems, especially LLMs.

Different in kind

Of course, the real social skills and knowledge are coming from the human. We're not yet ready to say that LLMs have social skills and social knowledge in the same robust sense that even three-year-old human beings have those skills and knowledge. But the machine is designed, or at least has emerged from its developmental process, in a way that exploits the fact that you will react to it as a social agent; and you, in turn, can exploit that fact about it.

Asymmetric quasi-social interactions, then, are interactions between a fully social agent and some partner – whether human, machine, or animal – that is not cognitively capable of full-fledged (social) joint action but that does respond in a way that productively invites further social responses from the social partner. There is a type of social exchange drawing upon the social skills of the fully social agent, even if the partner lacks anything like real social understanding.

Thus, quasi-sociality places relatively little cognitive demand on the junior partner. The junior partner needn't understand that the other is an agent, nor must the junior partner have beliefs, desires, or goals. The junior partner needn't intend to communicate or cooperate. The junior partner need not even be a conscious entity. Still, the junior partner must be more than a Roomba. It must be structured in such a way as to draw social behavior from the senior partner, reacting to the senior partner's social behavior in a way that solicits further social behavior and does so in a manner that importantly resembles social interactions as they transpire between two fully fledged social partners.

Resemblance, of course, is a matter of degree. Furthermore, it's a multidimensional matter of degree: Since social interchange is complex, there are multiple relevant dimensions of resemblance concerning agency and coordination, like exchanging social information or mindreading.

Quasi-sociality thus exists on a complex spectrum. This is exactly what we should expect, given a bird's eye view of the phenomenon, a view that doesn't myopically focus on adult humans as the only types of social partners. Complex social skills will, of course, not emerge in an instant – not developmentally in humans, nor phylogenetically in animal evolution, nor technologically in the design of AI systems. We should expect a wide range of quasi-sociality between entirely asocial Roombas and viruses, on the one extreme, and on the other extreme, fully social, explicitly cooperative, second-order attitude ascribing adult humans.

Towards quasi-social in-between phenomena

Let's discuss one more example that points to such interesting in-between phenomena. In Berkeley, there used to be simple delivery bots called Kiwibots, that would roll along sidewalks and onto campus to deliver small food orders (Chávez et al., 2023). These bots are mostly autonomous but require some remote human intervention, for example, when crossing intersections. The bots have digitally animated eyes, and they trigger nurturing behavior like Pleos do. Occasionally, of course, they wander off path or get stuck somewhere. Evidently, when this happens, passersby sometimes help the Kiwibots out. Maybe their cute and non-threatening appearance makes this more likely. We are not claiming that those Kiwibots are already in the realm of quasi-social partners. But one can easily imagine an update of the technology toward quasi-sociality. Suppose that when a Kiwibot gets stuck somewhere, it emits some mildly

distressed noise – "ooh, ooh" – and says, "Gosh, I'm stuck. Maybe someone will help me?" Suppose, then, that once it's in this situation, it can detect whether it has been helped. It can detect whether a person – a social agent and not another machine – has approached it, contacted it, and started it moving again along the desired path. After this, maybe it says, "Thank you so much for the help, friend!"

Such a design would be a small step along the path of robot quasi-sociality: an interaction pattern designed to productively trigger social behavior in the senior partner. If this interface proves useful so that more Kiwibots get to their destination faster, then one might imagine future versions with more sophisticated social interactions. For example, maybe people who order food can opt into letting the Kiwibot display their name and face. If the bot is delivering to a crowded room, or if the bot is not promptly unloaded, perhaps it can approach a bystander in a slow and seemingly timid way and scan the bystander's face for a friendly or welcoming expression. If the bystander's expression isn't classified as welcoming, the bot can terminate the interaction and maybe approach someone else. Upon detecting a face classified as welcoming, the bot might emit "Could you help me find Devan?" displaying a picture of Devan's face. "I have a delivery for them!"

In this way, one could imagine a progression of ever more sophisticated delivery bots, which ever more effectively exploit the social capacities of senior partners. Maybe at some point – who knows when? – they become genuinely conscious, genuinely capable of social emotion, and genuinely capable of knowing that you know that they know. But that's the end of the road. The quasi-sociality starts far before then. Along the way, we expect a wide, wide gray area of first quasi-sociality and then asymmetric sociality with the human as a senior partner.

Crossing over from quasi-sociality to minimal, though asymmetric sociality, requires a few more things from the junior partner. Among those things might be something like a justified ascription of minimal agency and minimal abilities of coordination, which include further capacities like social information exchange, mindreading, and commitment.

Observing the ongoing debates about the abilities artificial systems have, one can point to the fact that there are artificial systems that are only to a very small extent under our control. Those artificial systems seem to have some degree of "autonomy" and are even able to "adapt" and "learn". That is, some tools are already beginning to have agent-like properties, on a wide or liberal conception of agency (Strasser, 2022). We won't explore here exactly what such abilities consist of other than to remark that we are skeptical of bright lines.

Two Implications

We conclude with two implications. First, one striking feature of human asymmetric sociality and asymmetric joint action, especially between parents and infants or young children, is that the parents provide scaffolding for the child's developing sociality. By treating the child as a social partner, the parent helps *make* the child a social partner. When we are slightly aspirational in our interpretation of our children, reading into their actions and reactions maybe a little more sophistication than is really there, this helps those children rise into the roles and attributes we imagine for them. By trusting, we help make them trustworthy. By treating them as fair, moral, and sympathetic to others, we help make them more fair, moral, and sympathetic to others. This could potentially also be true for AI systems that are capable of learning from our interactions with them. Perhaps, for the right machines, if we treat them as social partners, this helps them develop the pattern of reactions that make them social partners.

Second, as the Replika and Kiwibot examples should make clear, there's potential for corporate exploitation. Our upgraded Kiwibot is innocent enough, but it is drawing upon the freely given goodwill of bystanders to achieve the corporate end of an efficient delivery. More problematically, if people really do fall in love with their Replika chatbots, then, of course, they will want to pay monthly fees to maintain the service, and they will pay extra for sexy pictures and they will pay extra for stylish clothes and fancy

features. There's obvious potential for lonely people to be exploited. Clever engineers of quasi-social AI systems could potentially become skilled at generating social reactions from users in a way that exploits human vulnerabilities for the sake of corporate interests. This could be especially the case if quasi-social AI systems are designed to generate real feelings of love and attachment. We don't want people committing suicide when their chatbot rejects them and we don't want someone diving out into traffic, risking their lives to save a Kiwibot from an oncoming truck.

We anticipate that the future will present humans with a diverse range of quasi-social and asymmetrical social AI partners. We have barely begun to think through the possibilities and implications.

References

Brandi, M.-L., Kaifel, D., Bolis, D., & Schilbach, L. (2019). The Interactive Self – A Review on Simulating Social Interactions to Understand the Mechanisms of Social Agency. *I-Com*, *18*(1), 17–31. https://doi.org/10.1515/icom-2018-0018

Bratman, M. E. (2014). *Shared Agency: A Planning Theory of Acting Together*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199897933.001.0001

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners*. https://doi.org/10.48550/arXiv.2005.14165

Brownell, C. A. (2011). Early Developments in Joint Action. *Review of Philosophy and Psychology*, *2*(2), 193–211. https://doi.org/10.1007/s13164-011-0056-1

Bryson, J. J. (2010). Robots Should be Slaves. In Y. Wilks (Ed.), *Natural Language Processing* (Vol. 8, pp. 63–74). John Benjamins Publishing Company. https://doi.org/10.1075/nlp.8.11bry

Butterfill, S. A., & Apperly, I. A. (2013). How to Construct a Minimal Theory of Mind. *Mind & Language*, *28*(5), 606–637. https://doi.org/10.1111/mila.12036

Carpenter, J. (2016). *Culture and Human-Robot Interaction in Militarized Spaces: A War Story*. Routledge.

Chávez, F., Pachón, S., & Rodriguez, D. (2023). *Kiwibot autonomous delivery robots, revolutionizing the future of robotic delivery.* https://www.kiwibot.com/

Coeckelbergh, M. (2012). *Growing Moral Relations*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137025968

Cole, S. (2023). "It's Hurting Like Hell": AI Companion Users Are In Crisis, Reporting Sudden Sexual Rejection. *Vice*. https://www.vice.com/en/article/y3py9j/ai-companion-replika-erotic-roleplay-updates

Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in Human–Robot Co-evolution. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00468

Darling, K. (2021a). *The new breed: What our history with animals reveals about our future with robots* (First edition). Henry Holt and Company.

Darling, K. (2021b). *Could You Kill a Robot Dinosaur?* Built In. https://builtin.com/robotics/kill-robot-dinosaur-the-new-breed-kate-darling-excerpt

Davidson, D. (1980). *Essays on Actions and Events*. Oxford University Press. https://doi.org/10.1093/0199246270.001.0001

Davidson, D. (1982). Rational Animals. *Dialectica*, *36*, 317–328. https://doi.org/10.1111/j.1746-8361.1982.tb01546.x

Davidson, D. (2001). *Subjective, Intersubjective, Objective: Philosophical Essays*. Oxford University Press.

Dennett, D. (1987). *The Intentional Stance*. The MIT Press.

Fletcher, L., & Carruthers, P. (2013). Behavior-reading versus mentalizing in animals. In J. Metcalfe & H. S. Terrace (Eds.), *Agency and joint attention* (pp. 82–99). Oxford University Press. https://books.google.com/books?hl=en&lr=&id=IjFpAgAAQBAJ&oi=fnd&pg=PA82&dq=info:ci1kW6u

6XlUJ:scholar.google.com&ots=9gdBzPxpkT&sig=dtgiTfS4v1B0FWDltDf3DVv0Xrg
Gunkel, D. J. (2012). *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. The MIT Press. https://doi.org/10.7551/mitpress/8975.001.0001
Gunkel, D. J. (2023). *Person, thing, robot: A moral and legal ontology for the 21st century and beyond*. The MIT Press.
Halina, M. (2015). There Is No Special Problem of Mindreading in Nonhuman Animals. *Philosophy of Science*, *82*(3), 473–490. https://doi.org/10.1086/681627
Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*, 243–259. https://doi.org/10.2307/1416950
Heyes, C. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences*, *21*(1), 101–114. https://doi.org/10.1017/S0140525X98000703
Heyes, C. (2014). False belief in infancy: A fresh look. *Developmental Science*, *17*(5), 647–659. https://doi.org/10.1111/desc.12148
Heyes, C. (2015). Animal mindreading: What's the problem? *Psychonomic Bulletin & Review*, *22*(2), 313–327. https://doi.org/10.3758/s13423-014-0704-4
Jensen, C. B., & Blok, A. (2013). Techno-animism in Japan: Shinto Cosmograms, Actor-network Theory, and the Enabling Powers of Non-human Agencies. *Theory, Culture & Society*, *30*(2), 84–115. https://doi.org/10.1177/0263276412456564
Lam, B. (2023, April 25). Love in Time of Replika. *Hi-Phi Nation*. https://hiphination.org/season-6-episodes/s6-episode-3-love-in-time-of-replika/
Michael, J., Sebanz, N., & Knoblich, G. (2016). The Sense of Commitment: A Minimal Approach. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01968
Murphy, M., & Templin, J. (2019). Replika: This app is trying to replicate you. *Quartz*. https://qz.com/1698337/replika-this-app-is-trying-to-replicate-you
Pacherie, E. (2013). Intentional joint agency: Shared intention lite. *Synthese*, *190*(10), 1817–1839. https://doi.org/10.1007/s11229-013-0263-7
Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512
Roose, K. (2023). Bing's A.I. Chat Reveals Its Feelings: 'I Want to Be Alive.' *The New York Times*. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html
Shevlin, H. (2021). *Uncanny believers: Chatbots, beliefs, and folk psychology.* https://henryshevlin.com/wp-content/uploads/2021/11/Uncanny-Believers.pdf
Silver, C. A., Tatler, B. W., Chakravarthi, R., & Timmermans, B. (2021). Social Agency as a continuum. *Psychonomic Bulletin & Review*, *28*(2), 434–453. https://doi.org/10.3758/s13423-020-01845-1
Singer, P. W. (2010). *Wired for war: The robotics revolution and conflict in the twenty-first century*. Penguin Books.
Strasser, A. (2020). From tools to social agents. *Rivista Italiana Di Filosofia Del Linguaggio*, *14*(2), Article 2. https://doi.org/10.4396/AISB201907
Strasser, A. (2022). From Tool Use to Social Interactions. In Janina Loh, Wulf Loh (Ed.), *Social Robotics and the Good Life: The Normative Side of Forming Emotional Bonds With Robots* (pp. 77–102). transcript Verlag.
Strasser, A. (2023). *On pitfalls (and advantages) of sophisticated large language models*. https://doi.org/10.48550/arXiv.2303.17511
Strawson, P. (1962). Freedom and Resentment. *Proceedings of the British Academy*, *48*, 187–211.
Trevarthen, C., & Aitken, K. J. (2001). Infant intersubjectivity: Research, theory, and clinical applications. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *42*(1), 3–48.
Vesper, C., Butterfill, S., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks*, *23*(8), 998–1003. https://doi.org/10.1016/j.neunet.2010.06.002