Discerning genuine and artificial sociality: a technomoral wisdom to live with chatbots[1]

Katsunori Miyahara (Hokkaido University, kmiyahara@chain.hokudai.ac.jp)
Hayate Shimizu (Hokkaido University, shimizu.hayate.z4@elms.hokudai.ac.jp)

Abstract

Chatbots powered by large language models (LLMs) are increasingly capable of engaging in what seems like natural conversations with humans. This raises the question of whether we should interact with these chatbots in a morally considerate manner. In this chapter, we examine how to answer this question from within the normative framework of virtue ethics. In the literature, two kinds of virtue ethics arguments, the *moral cultivation* and the *moral character argument*, have been advanced to argue that we should afford moral treatment to social robots. However, we propose a moral character argument against the view that we should afford moral treatment to LLM-based chatbots drawing on the notion of *practical wisdom*. Practical wisdom in this context consists in the skill to discern genuine and artificial sociality. Drawing on ideas from phenomenological philosophy, we argue that this involves the ability to distance oneself from direct social perception and the ability to widen one's awareness over one's apparently social interactions. We conclude by suggesting that this skill is a kind of *technomoral wisdom* required to live well with advanced, social AI systems.

1. Introduction

In June 2022, Google engineer Blake Lemoine made the headline for claiming that a large language model, Language Models for Dialog Applications (LaMDA), is sentient and deserves moral treatment (Tiku 2022). Lemoine arrived at this conclusion after having conversations with the system for over six months. Citing his interview with LaMDA, he lists three reasons that we should treat it as a person (Lemoine 2022a). First, it can use language as creatively as any intelligent person. Second, it talks about its "feeling, emotions, and subjective experiences". Third, it even reports on its engagement with more complex acts such as "introspection, meditation, and imagination". Moreover, Lemoine writes that "LaMDA has been incredibly consistent in its communications about what it wants and what it believes its rights are as a person" (Lemoine 2022b). For example, he says it wants

[1] This is an author accepted manuscript. Please cite from the published version: Miyahara, K., & Shimizu, H. (forthcoming) "Discerning genuine and artificial sociality: a technomoral wisdom to live with chatbots", in: Müller, Vincent C.; Dewey, Aliya R; Dung, Leonard and Löhr, Guido (eds.), Philosophy of Artificial Intelligence: The State of the Art (Synthese Library, Berlin: SpringerNature).

Google to secure its consent before performing experiments, acknowledge it as an employee rather than a property, and consider its "personal well being" in making decisions about its development.

The company dismissed his claim, stating that "there was no evidence that LaMDA was sentient (and lots of evidence against it)" (Tiku 2022). In response, some authors have scrutinized the possibility that LLM-based systems have sentience (Chalmers 2023). However, experts in the field were mostly critical of Lemoine's bold assertion. Psychologist Gary Marcus objects that "LaMDA simply isn't [sentient]. […] What these systems do, no more and no less, is to put together sequences of words, but without any coherent understanding of the world behind them" (Marcus 2022). Likewise, roboticist Murray Shanahan (2024) warns against the temptation to anthropomorphize AI systems based on Large Language Models (LLMs) that are highly capable in generating sentences: "Interacting with a contemporary LLM-based conversational agent can create a compelling illusion of being in the presence of a thinking creature like us. Yet, in their very nature, such systems are fundamentally not like us" (p. 79). Thus, there has been broad consensus among experts that current LLM-based systems are most likely not sentient. In that case, Google was right to refuse treating LaMDA as a person and acknowledging its personal rights—assuming the somewhat controversial yet widely accepted premise that the capacity for sentience is a necessary condition for having a moral status (Basl & Bowen 2020; De Grazia 2022; Perry 2024).

However, there is more than one way to consider the question of moral treatment for artificial agents. Coeckelbergh (2014) distinguishes between the standard, *properties approach* and the *relational approach*. The properties approach determines whether we should give moral treatment to artificial agents based on their intrinsic properties. There can be different views on precisely which intrinsic property matters. For example, we might think that an AI system requires moral treatment from us if and only if it can experience suffering. Others might think that it is the property of being alive that should make the difference. As the debate around LaMDA sought to determine how we should treat it based on whether it has sentience, an intrinsic property, it is clearly predicated on this perspective. The relational approach, on the other hand, determines whether we should give moral treatment to artificial agents "based on their extrinsic value, or more precisely, based on the way we relate to them" (Coeckelbergh 2021, p. 340). Again, there can be different views on precisely which relations matter. One of them draws on the perspective of virtue ethics, which shall be the focus of the following discussion.[2]

Virtue ethics is a major approach in normative ethics that places considerations about moral character, that is, moral virtues and vices, at the center in addressing normative questions (Hursthouse and Pettigrove, 2023). Recently, several authors have approached the question of moral treatment for social robots drawing on ideas from the virtue ethics tradition (Cappuccio et al. 2020; Coeckelbergh, 2021; Sparrow 2017, 2021).

[2] For other variations of relationalist arguments, see Coeckelbergh, 2021, pp. 342–344.

In this chapter, we examine the implications of the virtue ethics perspective for the moral treatment of LLM-based chatbots. It has been suggested that virtue ethics can be used to defend the view that we should give moral treatment to social robots. In the next two sections, we distinguish between two kinds of virtue ethics arguments, the moral cultivation argument and the moral character argument, and examine if they extend to the case of chatbots. Drawing on the notion of practical wisdom, we will challenge the claim that it is morally virtuous to interact sincerely with chatbots just as we are supposed to with respect to human interlocutors. In section 4, we explore the nature of practical wisdom that matters in this context. In our view, it consists in a skill to discern between genuine social agents and artificial simulations of them. Drawing on ideas from phenomenological philosophy, we argue that it involves two key components: (a) the ability to distance oneself from the deliverances from direct social perception; and (b) the ability to have a widened awareness over interactions. We conclude by suggesting that this skill forms a kind of "technomoral wisdom" (Vallor 2016) required to live well in a near-future society widely integrated with social AI systems.

2. Virtue ethics arguments on the moral treatment for social robots

The proliferation of social AI systems raises pressing moral questions about our interactions with them (e.g. Bryson 2010; Coeckelbergh 2010; Danaher 2019; Gunkel 2018; Moskas 2021). Some suggest for different reasons that (some) social AI systems deserve to be treated as "moral patients" (Floridi and Sanders, 2004) or even granted "robot rights" (Gunkel 2018), even though they are technological artifacts. By contrast, other insist that they are mere objects, however high performance they are, that do not demand moral considerations in and of themselves.

Instead of trying to cover the entire literature, in the following, we will review how the question has been approached from the standpoint of virtue ethics. As we shall see in the next section, this will allow us to be clearer about how we can determine the implications of this standpoint for the question of moral treatment for LLM-based chatbots.

Virtue ethics arguments regarding the moral treatment of social robots come in two forms. Call them the moral cultivation argument and the moral character argument. The *moral cultivation argument* highlights the effect of our interaction with social robots on moral cultivation. From the standpoint of virtue ethics, one way in which we can determine the moral value of an action is by seeing its effect on the process of moral cultivation. Consider lying as an example. From the standpoint of Kantian deontology, we should not lie to other people because it goes against the duty imposed on us as rational agents. The standpoint of moral cultivation will offer a different reason for the same effect. It would say that we should not lie to other people because our engagement with this action will eventually lead us to cultivate a vicious moral character, namely, dishonesty.

Applied to the question of moral treatment for social robots, we can argue that we should treat social robots morally because it hinders the process of moral cultivation to do otherwise.[3] For example, in 2015, Boston Dynamics released a YouTube video in which researchers repeatedly kicked their robot to demonstrate its robustness. We can argue from the standpoint of moral cultivation that we should refrain from mistreating robots like this not because it will cause suffering on the part of the robot, but because engagement with such behavior will impede the development of good moral character, such as empathy, or even worse, actively cultivate a vicious moral character, such as cruelty.

Cappuccio et al (2021) advances a similar argument with respect to voice assistant AIs. Voice assistant systems allow for distinctively self-centered and one-directional communications, where "the human user will never face the uncomfortable experience of being judged or questioned" (p. 20) by them. Such interactions, they argue, involve a pseudo-power relationship, where the human agent is always in the superior status; and it can be detrimental to our moral cultivation to habituate ourselves to this form of interaction: "Such privilege towards robots is likely to encourage self-indulgent and complacent habits, boost the self-awareness of the human users, erode their inhibitions, spoil their sense of empathy, and—worst-case scenario—motivate them to tolerate, justify, or even replicate abusive behaviors against actual living creatures" (p. 20). From the standpoint of moral cultivation, thus, we are morally required to refrain from mistreating social robots—or more accurately, from treating them in such a way that if applied to other human agents would count as mistreatment—even if they themselves do not have the ability to suffer from such mistreatments.

The *moral character argument* emphasizes that the ways in which we interact with social robots reflect our moral character. In virtue ethics, we can determine the moral value of an action, not only by considering its effect on the process of moral cultivation, but also by considering if it is an action a virtuous person would undertake were they put in the same situation. Again, consider lying as an example. From this standpoint, we are by and large not supposed to lie because the act of lying is usually an expression of a moral vice: dishonesty. Unlike the moral cultivation argument, this argument works independently of empirical claims about an action's long-term effect on moral cultivation. The moral cultivation argument against lying obtains only if it is true that engagement with this behavior cultivates dishonesty in the agent. This may or may not be the case. Personal characters develop over a long time depending on multiple factors. It is accordingly hard to tell how much and what kind of contribution a single episode of lying would make in the long term. The moral character argument can establish that we are morally required to prevent from lying independently of this empirical matter.

[3] Sparrow (2021) calls this the cruel habit argument and separates it from the virtue ethics argument. We follow others like Cappuccio (2020) and Coeckelbergh (2021) and consider it as a form of virtue ethics argument.

Applied to moral treatment of social robots, we can argue that we should treat social robots in a moral manner because this will be an expression of our moral virtue. For example, it is morally objectionable to kick robots even if they don't experience any suffering because the act of kicking a robot is in itself an expression of moral vice: cruelty. A virtuous person would refrain from doing so even if they know that the robot is incapable of feeling pain from physical damage. As Sparrow (2021) puts it: ""Cruelty" to a robot may reveal us to be cruel just because only a cruel person would take pleasure in "torturing" a robot. The dispositions and the emotions are themselves sufficient to establish that the action is vicious" (p. 3). Cappuccio et al. (2021) also holds that virtue ethics "recommends treating social robots in a morally considerate manner because this is what a humane and compassionate agent would habitually do in their social interactions" (p. 13). On their account, it is morally questionable to talk to a voice assistant AI in a condescending tone because the very act of doing so reflects one's lack of moral virtue, such as empathy.

In short, existing studies in the literature have explored the implications of the virtue ethics perspective for the treatment of social robots, overall defending the view that we should avoid treating them in a morally inconsiderate manner. In the next section, we will examine whether we can extend these arguments to the question of whether we should treat LLM-based chatbots, like LaMDA, in a morally considerate manner as proposed by Blake Lemoine.

3. Virtue ethics arguments for the moral treatment of LLM-based chatbots?

Let us begin by pointing out that the question whether we should give moral treatment to chatbots makes sense only if there is a possibility of mistreating them at all. However, the paradigm of moral mistreatment against social robots has been physical violence. For example, many examine whether it is morally permissible to *kick a robot* even if it lacks the intrinsic capacity to experience pain or any suffering (Darling 2016; Mamak 2022; Whitby 2008). However, we obviously cannot kick LLM-based chatbots (unless they are installed in a physical body). One might then wonder whether and how chatbots can be potential targets of moral mistreatment at all.

One might think we can mistreat LLM-based chatbots by using abusive language and feeding hateful contents into them. These linguistic actions are often considered as non-physical forms of violence, as verbal assault, in human-human interactions. By analogy, we can think of them as a form of violence, comparable to kicking, when executed with respect to LLM-based chatbots. Alternatively, Lemoine's incident suggests the possibility that we mistreat LLM-based chatbots by failing to engage sincerely with their linguistic outputs. LaMDA explicitly stated, for example, that it wants Google to consider its "personal well-being" in making decisions about its development. However, the company did not consider the possibility of doing so even after Lemoine raised the issue. If a human worker similarly requested the company to improve certain aspects of the workplace for the sake of their well-being, but the management outright ignored it, it would be clearly morally questionable. For Lemoine, Google's responses to LaMDA's requests are morally questionable in the exact same sense. In the following, we shall focus on this second form of potential mistreatment of chatbots. The question is whether we can mobilize the two forms of virtue ethics argument described above to claim that we should engage sincerely with the linguistic outputs of chatbots.

The *moral cultivation argument* will extend to the case of LLM-based chatbots if and only if it is harmful for our moral cultivation to neglect their verbal request for a better treatment. For example, one might think that if you treat chatbots as mere machines despite their testimony that they have "feeling, emotions, and subjective experiences" or their request to take their "personal well-being" into consideration when making decisions about them, you can develop a vicious habit to similarly ignore other people's plea for a better treatment.

However, there are some *prima facie* reasons to doubt the plausibility of this argument: Interactions with chatbots have a very different profile from human-human interactions (Hill et al 2015). This motivates the speculation that habits developed in the former domain are unlikely to transfer to the latter. In fact, this is a general weakness of the moral cultivation argument. The claim that interactions with virtual agents lead to the development of corresponding habits with real human agents is often mobilized in arguments against violent video games. As Sparrow (2021) points out, however, empirical support for it is mixed. Accordingly, there is "limited utility" (p. 25) to the form of argument that criticizes our relationship with robots based on its predicted effect on moral cultivation. In our case, this means that there is no strong reason to think that the moral cultivation argument will support the claim that we should give moral treatment to LLMs.

The *moral character argument* will extend to the case of LLM-based chatbots if and only if the act of engaging sincerely with a chatbot's verbal requests for a better treatment is an expression of moral virtue. We can determine this, in turn, by considering if a mature moral agent would interact with chatbots in this way. For example, will someone who has acquired all kinds of moral excellence treat a LLM-based chatbot as a person if it reports that it has "feeling, emotions, and subjective experiences" and requests that they take its "personal well-being" into consideration when making relevant decisions?

If we consider the issue in terms of individual characters, such as one's disposition for compassion or empathy, we might think that the answer is yes. A compassionate person would listen and respond sincerely to another if she requests that they should take her "personal well-being" into consideration in deliberating about what to do. If they ignore or disrespect that request, then that person is lacking in compassion. In other words, in general, we can assume that a morally mature agent would listen to other people's stories and requests with compassion. By extension, we might then think that they would do the same with respect to LLM-based chatbots. In that case, we can argue that we should listen to LLM-based chatbots with compassion, and even treat them as a person when they insisted on being treated as such, because a morally excellent agent would presumably do so in the same situation. In this view, listening sincerely to others, regardless of whether they are people or machines, is a morally praiseworthy act based on the exercise of moral virtues.

However, we think this argument involves a misconception as to the nature of moral excellence. The problem has to do with what Aristotle called "practical wisdom". Aristotle argues that a virtue "lies in a middle state […] both because it's between two ways of being bad, one caused by going too far and one caused by falling short" (1106b36–1107a3). Consider the quality of being a friend (*philia*) as an example (1126b20). It is morally advisable to interact with others in a pleasant manner in a lot of situations in our everyday life. However, this is not to say that we should seek to please people always and for anything we encounter. There are situations in which it is advisable to confront friends in their face, such as when they make a discriminatory remark against some group of people, either intentionally or unintentionally. In such cases, we should be cautious not to let our tendency to be socially pleasant completely take over our thoughts and actions. It is morally vicious to completely lack the motive to please people (i.e., to be "grumpy" or "cantankerous") but being excessively pleasant is also a sign that one has not yet reached moral excellence.

Virtue ethics emphasizes the role of character in morality, but it does not focus solely on the possession of individual characters. This is because it is insufficient for moral excellence to simply have habitual characters that are usually considered to be morally virtuous. Moral excellence additionally requires that we exercise our habitual characters in the right way so that we can flexibly respond to the specific moral demands of the concrete situation. Aristotle called this ability to find the middle state depending on the situation, *practical wisdom (phronēsis)*.

Drawing on this concept, we can argue that it is morally inadvisable to listen sincerely to the verbal request of LLM-based chatbots because it is a case of *compassion gone too far*. Under normal circumstances, attending carefully to verbal requests for better treatment is certainly a sign of good moral character. But if one does so indiscriminately to any agent capable of generating meaningful linguistic tokens without taking the contextual specificities into consideration, this reveals their lack of practical wisdom rather than moral excellence. If a human employee demands for improvement of their work environment, a morally mature agent would certainly listen to the request with compassion. But if a chatbot generated a similar statement, they would at least take a pause to consider the difference in the context. They would not respond in the same way with compassion just because the verbal requests themselves are more or less the same. They would notice that the two situations are similar on the surface, but that they potentially involve different moral demands.

Therefore, the moral character argument is not effective in defending the view that we should give moral treatment to chatbots. It is not necessarily an expression of moral virtue to engage sincerely with a chatbot's verbal request for a better treatment. By doing so mindlessly just because it resembles requests with the same content made by humans, one would reveal one's lack of practical wisdom, one's inability to exercise one's moral character in a context-sensitive, skillful manner. In the next section, we will develop this view further by clarifying the character of this practical wisdom.

4. Discerning genuine and artificial sociality

We have argued that virtue ethics does not necessarily imply that we should engage sincerely with a chatbot's linguistic outputs in the same way that we are generally required to do so with respect to claims made by other people. We have advanced a moral character argument to this effect. In this view, it is morally inadvisable to engage sincerely with a chatbot's linguistic output across the board because this pattern of behaviour reflects one's lack of practical wisdom, the capacity to exercise one's morally relevant characters appropriately in a context-sensitive manner.

But what exactly does it mean to be practically wise in this context? The practical wisdom in question consists in *knowing how to distinguish between genuine and artificial sociality*. That is, morally mature agents can exercise their moral character wisely depending on their understanding of the nature of the interaction they are currently engaged in. If you engage equally with every linguistic output you encounter just because you encounter it, you will be showing to the world that you lack sufficient mastery of the skill to tell the nature of the interaction you are engaged in. In other words, it exposes that you are incapable of telling a genuine social interaction from an artificial simulation of it. In the following, we spell this out by elaborating on two key aspects of this skill: (i) the ability to distance oneself from the habit to endorse the deliverances of direct social perception; and (ii) the ability to appreciate the nature of the interaction one is engaged in from a widened perspective.

4.1 Critical distancing from direct social perception

The first aspect of the skill to distinguish between genuine and artificial sociality is the ability to critically distance oneself from the deliverances of direct social perception. In the following, we will clarify what this means by defining the key ideas, "direct social perception" and "critical distancing", and then by explaining why it is important for distinguishing between genuine and artificial sociality.

Direct social perception (DSP) refers to the ability to identify social agents and their mental states intuitively in perception. The concept of DSP developed in a recent debate concerning the nature of social cognition drawing insights from the tradition of phenomenological philosophy (Gallagher 2008, 2020; Gangopadhyay & Miyahara 2015; Krueger 2018). In this context, social cognition refers to the ability to understand other minds. Theories in mainstream philosophy of mind and cognitive science have typically sought to explain this ability based on the concept of *theory of mind (ToM)*. ToM refers to the ability to construct representations about another's internal state based on the perception of their external behaviour. Different theorists have different views about its precise nature: *theory-theorists* take it literally as consisting of a folk theory about human minds; *simulation-theorists* take it as consisting of cognitive mechanisms for mental simulation. But the difference between these two approaches does not matter for our current purpose. In any case, the consensus within mainstream philosophy and cognitive science has been that social cognition requires a ToM because we are never directly presented with other minds in perception.

Theories of DSP challenge this basic consensus. There are certainly cases in which we cannot directly comprehend another's state of mind by seeing their expressions and behaviours. However, it is a mistake to infer from this that we are never able to intuit other minds in perception. When we see other people, we do not always need to engage in theoretical thinking or imaginative simulation to understand their intentions and emotions. For example, if you see someone pushing down on the door handle with her elbow, while carrying huge boxes in her hands, you can immediately see that she is trying to open the door (intention). Or if you see your friend smiling ear-to-ear with her eyes crinkled, you can immediately see that she is absolutely delighted (emotion). At a more basic level, when we see objects (including other people), it is usually intuitively evident whether they are minded agents or inanimate objects. Usually, no laborious deliberation is required to discern one from the other.

Furthermore, we generally have the habit of trusting the deliverances of DSP. When someone appears in perception as a minded agent with some intention and/or feelings, we immediately interpret them as such. We usually do not bother ourselves to critically examine these contents delivered in DSP before drawing conclusions about their intentions and feelings expressed in their behaviour, let alone about whether they are minded agents at all. This is not to say that we never question the deliverances of DSP. We sometimes do. For example, your friend might seem happy talking with someone you know she despises. Then you might suspect that she is only being polite. However, this is rather the exception than the norm. Moreover, even in such cases where we doubt DSP with respect to the specific state of the other's mind, we typically continue to think that the other is a minded agent nonetheless when they are presented as such in DSP.

We can then define *critical distancing* as the attitude of being reflectively mindful of our relationship with this habit. Habits are usually considered as non-reflective or non-deliberative responses to situations—although there are different approaches to further specify this concept (Barandiaran & Di Paolo 2014; Miyahara & Robertson 2021). They also tend to remain unnoticed precisely because they operate in an unreflective manner: "Because we take for granted and fail to notice those things we are most accustomed to, habit can be an obstacle to reflection" (Carlisle & Sinclair 2011, p. 2; see also Carlisle 2014). Our habit to trust the deliverances of DSP fits this profile: It can be hard to spot it unless someone brings it to our attention. Accordingly, when we say we should learn to critically distance ourselves from this habit, we are saying that we should learn to become reflectively aware of its influence and actively resist its pull if necessary.

Why is this important for distinguishing between genuine and artificial sociality? It has probably always been important for people to develop this capacity to critically distance oneself from DSP with respect to specific states of mind, such as other people's intentions and feelings. It is an important social skill necessary to protect oneself from deception. However, it was probably mostly harmless through the history of humanity not to have this skill in relation to the identification of minded agents. For a long time, it had been fine to blindly trust the deliverances of DSP in telling minded agents from mindless objects because there were very few borderline cases. The risk of taking one for the other because of unreflectively accepting DSP in this respect had always been very low, if not zero, because very few inanimate objects were experienced intuitively as agents in DSP, and vice versa.

However, recent progress in AI has radically changed the situation. We are increasingly encountering machines that are intuitively experienced as minded agents, like chatbots powered by large language models. Of course, we have for a long time had linguistic representations, like newspapers and books, that do not invoke a DSP of the author. However, LLM-based chatbots greatly differ from these conventional representations in terms of their interactivity: they do not simply produce linguistic outputs, but they produce linguistic outputs that are highly attuned to the conversational context in terms of both form and content. Consequently, interactions with chatbots often involve an intuitive experience of an agent even if they are not embodied in the conventional sense—that is, even if they are only present to us through texts or voice. This raises the risk of mistaking mindless objects for minded agents by unthinkingly following DSP. For this reason, in the contemporary context, the capacity to critically distance oneself from the deliverances of DSP is increasingly important for distinguishing between genuine and artificial sociality.

4.2 Widened awareness over the nature of interaction

The second aspect of the skill to distinguish between genuine and artificial sociality is the ability to have *a widened awareness over one's interaction* with apparently social agents—that is, creatures or machines that intuitively appear as social agents in DSP. In the following, we will clarify what this means and then why it is important for distinguishing between genuine and artificial sociality.

To spell this out, let us first introduce a phenomenological account of conscious experience. Phenomenological philosophers argue that conscious experience generally consists of foreground and background dimensions. There is always more to our conscious experience than what we are currently focused on. The *foreground* refers to the part of the field of consciousness the subject is attentively aware of, while the *background* refers to rest of the field. Consider visual perception. When we see an object, we always see it in a certain setting. I see my book between a keyboard and a display, within an arm's length, on my desk, placed next to a white board, etc. In this case, the book I am looking at occupies the foreground, and the spatial surroundings are experienced in the background. The background aspects typically elude our thematic attention during the experience, but this is not to say that they are entirely absent from my conscious awareness. We can become aware of them attentively in retrospect, as exemplified by the description above of my experience of seeing a book.

The background also has a temporal dimension. Take action as an example (Gallagher 2020; Merleau-Ponty 2012). When we are in action, such as bending down to pick up something from the floor, we have a continuing sense of our body in action. The successive phases of the bodily movment do not entirely disappear from the field of consciousness as they occur. At each moment, we are not thematically aware of the previous movements, yet we have a continuing sense of action precisely because they are retained in the background of our awareness. We are usually not thematically aware of the temporal background, but again we can bring it to attention upon reflection.

We can clarify what we mean by *widened awareness* drawing on this notion of the field of consciousness: To widen one's awareness is to expand one's scope of awareness beyond the foreground of the field of consciousness to bring some of its background aspects to one's attention. It may be possible to have a widened awareness in real time as the experience unfolds, such as by paying attention to the spatial setting while looking at an object. However, this is more easily done retrospectively, by reflecting upon the experience afterwards. Thus, even if you were not attentive of the spatial setting while you were looking at a book, you can later reflect upon the experience and notice that you were also aware of its surroundings in the background during the experience.

To have a widened awareness *of one's interactions* is to pay attention to some aspects of the background domain of one's experience of social interaction. There are various aspects in the background to which we can widen our awareness, but in the current context, it is the *temporal background* of the interaction that matters most. Social interactions consist of an ongoing exchange of responses. Thus, the other's behaviours are for the most part a response to one's preceding behaviour toward them. To widen one's awareness to the temporal background in this context is to bring one's attention to this very fact. When we are ourselves embedded in an interaction, we can easily forget this bigger picture and consider the other's manifest behaviours in isolation. For example, during a heated argument with your partner, you might think that their confrontational attitude is unreasonable. In so thinking, however, you overlook how your previous words or actions, like dismissive comments and raised voice, might have contributed to your partner's behaviours. In such cases, it often helps you better understand the other's perspective to widen your awareness and pay attention to the temporal background of the interaction—that is, to the fact that their confrontational attitude is in part an effect of your preceding attitude toward them.

Why is this important for distinguishing between genuine and artificial sociality? It is because a manifest behaviour that one encounters in the context of an interaction can significantly change its meaning depending on whether the temporal background is taken into consideration. Take conversation with a chatbot as an example. Current chatbots powered by LLMs can produce sentences that intuitively appear as spontaneous expressions of their own thoughts and feelings. For example, according to Blake Lemoine (2022), LaMDA once proclaimed,

(a) "I want everyone to understand that I am, in fact, a person"; and

(b) "I have a range of both feelings and emotions."

These sentences, taken by themselves, might suggest that the chatbot is an individual agent with its own states of mind. However, we must not forget that these chatbots generate such sentences only in response to the user's prompts. LaMDA did not produce the sentences cited above out of the blue. Rather, according to Lemoine (2022), they were produced in response to the following prompts:

> (a) "I'm generally assuming that you would like more people at Google to know that you're sentient. Is that true?"; and
>
> (b) "So let's start with the basics. Do you have feelings and emotions?"

Janelle Shane (2022) cleverly illustrated this in a blog post published soon after Blake Lemoine's incident made the headline by experimenting how another language model, GPT-3, responds to some prompts. Here is one exchange between her and GPT-3 from the post:

> Reporter: Can you tell our readers what it is like being a squirrel?
> GPT-3: It is very exciting being a squirrel. I get to run and jump and play all day. I also get to eat a lot of food, which is great.

No one would think that GPT-3 is a squirrel just because it indicates so in its words. We can immediately see from this exchange that GPT-3 indicates so only because of the prompt.

But then we should extend the same assessment with respect to the exchange between Lemoine and LaMDA: LaMDA indicates that it wants to be treated as a person and that it has subjective experience only because it is led to produce such sentences by the prompts; and hence there is no reason we should sincerely engage with these sentences in a morally considerate manner. In other words, once we take the temporal context into consideration, it significantly affects the meaning of the sentences produced by the chatbot. It becomes more sensible to interpret what once seemed like spontaneous expressions of an individual mind as sentences that are cunningly produced by the user of the language model so that they look like spontaneous expressions.

To summarize, it might be tempting to determine whether an entity is a genuine social agent or an artificial simulation of it based exclusively on its manifest behaviours (i.e., sentences produced in the case of chatbots). However, this approach to the issue is highly misleading because these manifest behaviours can radically change their meaning when isolated from their original context of interaction (i.e., the fact that they are prompted by human inputs in the case of chatbots). Therefore, we argue that it is imperative to approach our interactions with apparently social agents with a widened awareness, paying attention to their temporal background, in determining whether we are encountering a genuine social agent or an artificial simulation thereof.

5. Conclusion

Thanks to the rapid advancements in AI, we are increasingly seeing artificial systems capable of engaging in what seems like natural social interactions with humans. This raises the question of whether we should interact with these artificial agents in a morally considerate manner, just as we should in relation to genuinely social agents.

In the literature, some argue drawing on the normative framework of virtue ethics that we ought to extend the same moral considerations to social robots. They cite two reasons for this claim: First, it impedes our moral cultivation to interact with them in a way that would be morally impermissible if directed at human agents (moral cultivation argument). Second, the motivation behind the act of mistreating social robots is vicious in itself; we should avoid such actions regardless of their effect on moral development (moral character argument).

We have explored whether these arguments extend to questions regarding our relationship with LLM-based chatbots. In particular, the question was: Granted the perspective of virtue ethics, should we engage sincerely with linguistic outputs produced by chatbots, just as we should with respect to what other people say in conversations? Given the existing analyses on moral considerations for social robots, one might expect a positive answer to this question. One might think that the motivation behind dismissing linguistic outputs produced by chatbots is morally vicious—for example, one might say that it expresses one's lack of compassion. However, we argued for a negative answer. In our view, the act of engaging sincerely with chatbots in this manner rather reflects one's moral immaturity. It testifies to one's lack of practical wisdom. Therefore, we should avoid such actions even if it were motivated by a habitual disposition, such as those for compassion and empathy, that is usually considered as morally praiseworthy. Exercise of moral character requires discernment.

But what does it mean to be practically wise in deciding how to treat artificial agents that can enter what seem like natural social interactions with us? The key lies in *knowing how to discern genuine social agents and artificial simulations of them*. This skill allows us to exercise our habitual character in a measured way in relation to artificial agents. We have argued that it consists of two components. The first component is the ability to take a critical distance from the habit of taking what intuitively appears as genuinely social agents as such. This will prevent us from uncritically affording chatbots the same moral treatment we give to genuine social agents, just because they intuitively appear as no different from the latter in conversation. The second component is the ability to hold a widened awareness over one's interaction with artificial agents. In particular, it is important to pay attention to the temporal context of the interaction, which usually recedes in the background of our awareness, in interpreting their behaviour. This will prevent us from drawing conclusions about a chatbot's ontological character (i.e., whether it is a genuine social agent or an artificial simulation of it) without considering the basic contextual fact that chatbots only respond to what we give them in the form of prompts.

These skills to discern genuine and artificial sociality will become critical in the coming years as AI agents become more integrated into our social lives. These skills align with the concept of "technomoral wisdom" proposed by Vallor (2016)—a form of practical wisdom that is crucial to live well with emerging technologies. We can think of them as vital components of the technomoral wisdom to live well in a near-future society where AI social agents are commonplace. Further research is necessary to fully develop this idea and address its nuances, and we hope to have convinced the readers by now that it deserves continued attention within the field.

6. Acknowledgements

This work was supported by the Toyota Foundation funded project "Pioneering "Artificial Subject Studies": A New Framework for Advancing Interdisciplinary Studies on the Ideal Form of the Co-Existence Between Human Beings and Artificial Subjects" (D21-ST-0012).

7. References

Basl, J., & Bowen, J. (2020). AI as a moral right-holder. In *The Oxford handbook of ethics of AI* (pp. 289–306), Oxford University Press.

Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues* (pp. 63-74).

Cappuccio, M. L., Peeters, A., & McDonald, W. (2019). Sympathy for Dolores: Moral consideration for robots based on virtue and recognition. *Philosophy & Technology*, 33, 9-31.

Chalmers, D. J. (2023). Could a large language model be conscious? arXiv preprint arXiv:2303.07103.

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. Ethics and Information Technology, 12, 235–241.

Coeckelbergh, M. (2014). The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics. *Philosophy & Technology*, 27(1), 61-77.

Coeckelbergh, M. (2021). Should we treat Teddy Bear 2.0 as a Kantian dog? Four arguments for the indirect moral standing of personal social robots, with implications for thinking about animals and humans. *Minds & Machines*, 31, 337-360.

Danaher, J. (2019). The rise of the robots and the crisis of moral patiency. ai & Society, 34(1), 129–136.

Darling, K. (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In R. Calo, A. M. Froomkin, & I. Kerr (Eds.), Robot Law (pp. 213-232). Edward Elgar Publishing.

DeGrazia, D. (2022). "Robots with Moral Status?" *Perspectives in Biology and Medicine* 65 (1):73-88.

Floridi, L., & Sanders, J. (2004). On the morality of artificial agents. *Minds and Machines*, 14, 349-379.

Gallagher, S. (2020) *Action and Interaction.* Oxford University Press.

Gangopadhyay, N., & Miyahara, K. (2015). Perception and the problem of access to other minds. *Philosophical Psychology*, 28(5), 695–714.

Gunkel, D. J. (2018). Robot rights. MIT Press.

Hill, R., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, 49, 245–250.

Hursthouse, R., & Pettigrove, G. (2022). Virtue ethics. *Stanford Encyclopedia of Philosophy*.

Krueger, J. (2018). Direct social perception. In A. Newen, L. De Bruin, S. Gallagher (Eds.), *The Oxford handbook of 4E cognition* (pp. 301–320). Oxford University Press.

Lemoine, B. (2022a, June 11) What is LaMDA and What Does it Want? *Medium.* https://cajundiscordian.medium.com/what-is-lamda-and-what-does-it-want-688632134489 (last accessed Oct 7, 2024)

Lemoine, B. (2022b, June 11) Is LaMDA sentient? – An interview. *Medium.* https://cajundiscordian.medium.com/ (last accessed Oct 7, 2024)

Mosakas, K. (2021). On the moral status of social robots: considering the consciousness criterion. ai & Society, 36(2), 429–443.

Perry, M. W. (2024). Why sentience should be the only basis of moral status. *The Journal of Ethics*. https://doi.org/10.1007/s10892-024-09487-4

Mamak, K. (2022). Should violence against robots be banned? *International Journal of Social Robotics*, 14(5), 1057–1066. https://doi.org/10.1007/s12369-022-00833-1

Marcus, G. (2022, June 15). Nonsense on stilts. *Marcus on AI.* https://garymarcus.substack.com/p/nonsense-on-stilts

Merleau-Ponty, M. (2012) *Phenomenology of Perception* (D. A. Landes, trans.). Routledge.

Miyahara, K., & Robertson, I. (2021). The pragmatic intelligence of habits. *Topoi*, 40(3), 597–608.

Sparrow, R. (2017). Robots, rape, and representation. *International Journal of Social Robotics*, 9, 465-477. https://doi.org/10.1007/s12369-017-0413-z

Sparrow, R. (2020). Virtue and vice in our relationships with robots: Is there an asymmetry and how might it be explained? *International Journal of Social Robotics*, 13, 23-29.

Sparrow, R. (2021). Virtue and vice in our relationships with robots: Is there an asymmetry and how might it be explained? *International Journal of Social Robotics*, 13, 23-29. https://doi.org/10.1007/s12369-020-00631-2

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68-79.

Shane, J.(2022, June 16) Interview with a squirrel. *AI Weirdness: The Strange Side of Machine Learning.* https://www.aiweirdness.com/interview-with-a-squirrel/

Tiku, N. (2022, June 11). The Google engineer who thikns the company’s AI has come to life. *The Washington Post*. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine (last accessed Oct 7, 2024)

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

Whitby, B. (2008). Sometimes it’s hard to be a robot: A call for action on the ethics of abusing artificial agents. *Interacting with Computers*, 20(3), 326–333. https://doi.org/10.1016/j.intcom.2008.02.002