Digital Society (2024) 3:2
https://doi.org/10.1007/s44206-023-00086-8

ORIGINAL PAPER

Social Agency for Artifacts: Chatbots and the Ethics of Artificial Intelligence

John Symons[1] · Syed Abumusab[1]

Received: 20 April 2023 / Accepted: 27 November 2023

Abstract

Ethically significant consequences of artificially intelligent artifacts will stem from their effects on existing social relations. Artifacts will serve in a variety of socially important roles—as personal companions, in the service of elderly and infirm people, in commercial, educational, and other socially sensitive contexts. The inevitable disruptions that these technologies will cause to social norms, institutions, and communities warrant careful consideration. As we begin to assess these effects, reflection on degrees and kinds of social agency will be required to make properly informed decisions concerning the deployment of artificially intelligent artifacts in important settings. The social agency of these systems is unlike a human social agency, and this paper provides a methodological framework that is more suited for inquiry into artificial social agents than conventional philosophical treatments of the concept of agency. Separate aspects and dimensions of agency can be studied without assuming that agency must always look like adult human agency. This revised approach to the agency of artifacts is conducive to progress in the topics studied by AI ethics.

Keywords Artificial intelligence · Artificially intelligent social agents · Sexbot · Chatbot · Social Agency · Large Language Models

1 Introduction

Artificially intelligent systems are already part of our social world, appearing in commercial settings, in social media, in entertainment, and increasingly in caring roles in medical settings, as companions and as intimate partners (See De Gennaro et al., 2020; Gillath et al., 2023). AI and other data science technologies influence the social aspects of our lives in ways that are beginning to attract careful empirical and theoretical consideration (Natale, 2021). Ethical judgments

✉ John Symons
johnsymons@ku.edu

[1] Department of Philosophy, Center for Cyber Social Dynamics, University of Kansas, Lawrence, KS, USA

Published online: 27 December 2023

and policy-making decisions concerning the deployment of AI systems ought to consider not only their harmful effects on individuals but should also take into account their potential to result in socially harmful consequences (Symons and Elmer, 2022). To understand the social harms that can result from AI, it will be necessary to understand the ways in which the actions of AI systems can transform communities, social norms, and institutions.

AI obviously operates in social contexts, it can do so with some degree of autonomy, and it can manifest some degree of adaptability and goal-directed behavior, but its *actions* certainly do not take the same form as those of a human adult. Because of this difference, it is tempting for philosophers to simply deny that AI systems can count as agents at all. For many philosophers, the idea of treating artifacts as agents is a straightforward conceptual confusion or a category mistake. Technologists, by contrast, generally err in the other direction. They are tempted to assume a broad behaviorist conception of what both people and artifacts do, thereby denying a conceptually meaningful distinction between what a human person does and what an artifact does. Thus, for example, from Sam Altman's perspective, calling a large language model (LLM) a *stochastic parrot* is not necessarily an insult since in his view *we* are also nothing more than stochastic parrots.[1] With this move, the tech enthusiast circumvents the problem of AI agency by denying that there is anything special or especially puzzling about human agency.

While technologists blur the distinction between the activity of people and artifacts, traditional philosophers tend to focus on the kind of agency exhibited by normal adult human persons. This human-centered focus can also be an obstacle to clear reflection in this domain. As we shall see, to make sense of the social agency of artificially intelligent systems it is as important to avoid cleaving too closely to traditional philosophical conceptions of agency as it is to avoid simple-minded behaviorism. Both approaches to agency, we argue, pose an obstacle to the ethical evaluation of many of the most socially significant near-future uses of AI.

2 Why Does a Philosophical Analysis of the Social Agency of AI Matter?

Many of the concerns about social harms that motivate relatively recent science fiction are now pressing practical realities. For example, it is already the case that versions of the relationship portrayed in Spike Jonze's, 2013 movie *Her* are now, at least partially, realized. Millions of people interact with AI companions like *Replika*, and many consider those relationships deeply meaningful and satisfying (Weber-Guskar, 2022). It is currently unclear how we should think about the moral implications of these relationships.

[1] Sam Altman, head of OpenAI recently tweeted "i am a stochastic parrot, and so r u". https://twitter.com/sama/status/1599471830255177728?lang=en (Dec 22, 2022).

2.1 CarynAI, Duet, and the Problem of Socially Disruptive Technologies

Consider, for example, the *Snapchat* influencer Caryn Marjorie, who markets an AI-driven interactive representation of herself, CarynAI, trained on her video and audio recordings. This bot is marketed as providing romantic companionship to customers and she describes the chatbot version of herself as a “virtual girlfriend.” Customers pay $1/min for the service, and as of May 2023, she had around 1000 clients and anticipates eventually having around 20,000 of her 1.8 million Snapchat followers as customers (Sternlicht, 2023). Initial assessments of such a chatbot service from within current AI ethics are likely to focus on topics such as Ms. Marjorie’s proprietary rights to her representation and the ethics of consent, safety, and privacy. Most AI ethicists (at least within the Anglo-American tradition) generally embrace broadly individualistic and expressivist moral ideals and are less inclined to consider disruptions to social institutions and relations as harms that should be given moral consideration.[2] There are exceptions. For example, AI ethicists who are informed by feminist philosophy are likely to move beyond an individualistic perspective to examine some social implications of technologies like these. They might, for example, consider the commoditization of a “virtual girlfriend” and the role of gender in chatbot technologies. While feminist scholars are well equipped conceptually to lead the consideration of the implications of these technologies for existing social relations and institutions, at present, the implications of non-human agents intervening in human social systems, in particular the implications of these artifacts entering into close personal relationships with us, are poorly understood and understudied. Even if the human owners and customers of such systems are optimizing individual utility, have their privacy protected, and fully understand and consent to the service (a big *if*, of course), there remain broader questions of social harm and damage to valuable social institutions. Currently, it is not clear how we should think about the ethics of technologies that intervene in social relations in ways that harm social norms and institutions (see Symons and Elmer (2022)).

Consider another example: Google Meet (a videoconference service provided by Google) will soon deploy an AI system *Duet* that can attend meetings on behalf of users to deliver key talking points, summarize and take notes of meetings on behalf of users, and provide users who arrive to meetings late a private summary of the action that they missed.[3] Using AI as a proxy in these contexts can seem like an attractive and convenient option on an individual level. However, the social dynamics of meetings will change fundamentally under these conditions. Will one’s colleagues tolerate talking to one’s AI proxy instead of oneself? What if everyone sends their AI proxies? It is likely, for instance, that if this technology is widely adopted, meetings as we have known them will no longer take place. Perhaps, few of us will mourn the death of the old-fashioned business meeting, but such disruptions have unforeseen consequences.

[2] There is increased interest in Confucian approaches to these questions, see for example Zhu, 2020 that engages with the effects of technology on social roles as traditionally conceived in Chinese thought.

[3] https://workspace.google.com/blog/product-announcements/duet-ai-in-workspace-now-available (last accessed August 29 2023).

The nature of these implications will depend in part on the degree and kind of agency exhibited by these artifacts. Thus, to morally evaluate these and related systems, we must first understand the ways in which they act in the social world.

2.2 Unpacking Social Agency

Agency has been studied extensively in the philosophical literature (Schlosser, 2019). At its core, the concept of agency has been understood to capture the capacity of individuals to act intentionally and to have control over their effects on their environments in an adaptive manner. Agency has been explored in a variety of philosophical subdisciplines and traditions and has been a central topic in recent ethics, metaphysics, epistemology, and philosophy of mind. Discussions of AI and agency in the philosophical literature are extremely diverse, and we will not be able to engage with them all at length in this paper (but see Swanepoel (2021) for an excellent overview).

We argue first that agency can take a variety of forms and that it exists in differing degrees in a variety of contexts. In Sect. 5, we provide reasons to believe that the AI ethics community ought to set aside what we call the *threshold model of agency* that has been dominant in neighboring subdisciplines. In addition to arguing that we ought to distinguish dimensions and degrees of agency on empirical grounds, we also reject the threshold account on methodological grounds; as we explain below, a philosophical strategy that insists on a single sharp criterion or set of criteria that marks the threshold between agents and non-agents is an obstacle to carefully deliberate about questions in the ethics of technology. Instead, understanding aspects of agency and recognizing that they can be productively studied in terms of dimensions and degrees are both realistic and more methodologically fruitful in the ethics of AI than traditional threshold accounts.

In addition to criticizing views of agency drawn from recent analytic philosophy, we are also critical of ordinary pre-theoretical or common-sense capacities to judge what does and does not count as an agent. In Sect. 7, we show how ordinary common-sense judgments with respect to agency are simply too coarse-grained, vague, and misleading to allow careful and accurate moral judgments concerning artificially intelligent social agents. This latter point is relatively straightforward to introduce: It is already the case that we are sometimes deceived into thinking that an AI conversation partner or a robot is a human person. Given unregulated market pressure, corporate producers of AI technologies are likely to generate increasingly deceptive and manipulative products that fool us in ways that become difficult for us to detect. With the development of increasingly sophisticated techniques that exploit vulnerabilities in human social psychology, common sense alone will have limited usefulness. In this scenario, we will no longer trust our initial gut reactions or intuitive assessments of agency.[4]

[4] A folk psychological conception of agency detection along the lines Dennett describes in *The Intentional Stance* (Dennett, 1987) will be of little assistance in cases where we find ourselves devoting energy and attention to determining the nature of the beings with whom we are talking and interacting. The challenge here is that unaided common sense is not equipped to detect agent behavior in suitably sophisticated AI.

AI systems can possess some aspects of agency—they can obviously be considered agent-like to some degree and in certain domains of action. But it is also the case that one can productively analyze and evaluate agency in artifacts without assuming that artifacts have the same capacities or moral standing as typical adult humans.

We will argue, specifically, that some AI systems should be understood as *social* agents. They should be understood as responding to social norms, institutions, and structures, and they can intervene autonomously in ways that affect human social interactions and hierarchies. The social agency of AI is where core critical ethical questions emerge with respect to their effects on close personal relationships, economic transactions, social hierarchies, social institutions, and political life. AI ethicists will be required to understand new kinds of harm that result from these developments. Social harms—harms to social institutions, relationships, norms, and the like—are not always directly reducible to distributive harms to individuals or groups (Symons and Alvarado, 2022). Philosophers are beginning to recognize the deleterious effects of many technological developments on social trust, on the quality of our political discourse, on valuable cultural institutions, and on the quality of our communities. As corporations begin to deploy AI with social agency, we can preempt at least some of the social harms that may result and cultivate an ethos where we aim to ameliorate rather than degrade and devalue our social and cultural environments.

3 How the Social Agency of Artifacts Changes Our Social Lives: Linguistic and Non-linguistic Social Interventions

The effects of AI on social aspects of our lives are likely to be highly consequential, at least as consequential as Internet technologies and social media platforms have been over the past three decades. Internet technologies are widely suspected to have caused degraded levels of social trust and to have harmed the quality of close personal relationships, social norms, and institutions (Gao et al., 2020; Symons & Elmer, 2022). While there is some disagreement with respect to the causal role of Internet technologies and social media platforms in a range of social ills, the emergence of AI systems that act directly in ways that change our social lives should add urgency to reflection on social harms.

It is important to begin thinking in systematic philosophical ways about how and whether carebots will change our sense of responsibility toward the elderly and infirm, whether sex robots and “virtual girlfriends/boyfriends” will change marriage and dating norms and practices, how artificial conversational partners will change our view of friendship, how bots that we train on our own data and send to work on our behalf will change economic behavior, etc. Friendships, filial, parental, spousal, and other close personal relationships are typically thought to have special status, and they are each associated with distinctive kinds of value and obligation. If we regard these relationships as important, then we are likely to think that the harms that might result from AI require us to think carefully about the degrees and kinds of social agency exhibited by the artifacts deployed in our social environments.[5]

3.1 Embodied Social Agency: Sex Robots and Carebots

At this point, in the development of social AI technologies, it is easiest to focus on chatbots as the core manifestation of the social agency of artifacts. Nevertheless, we recognize that linguistic interaction is not the only way that an artifact can play a role as a social agent.[6] Embodied AI systems that act non-linguistically in social ways in the physical world also involve important and distinctive questions. Sex robots, for instance, have socially significant features; most obviously, their bodies have social significance insofar as they *refer* to the age, gender, and racial characteristics of human beings (Ma et al., 2022; Van Grunsven, 2022; Jecker, 2023).[7]

The design of the bodies of sex robots and the ways in which they act will be shaped by market demand. This demand is shaped by existing social conditions with all its familiar gender, racial, class, and age markers and hierarchies. While existing social vices will be manifest and perhaps reinforced by sex robots in the short term, technological innovations will also shape the preferences of consumers insofar as new possibilities and new kinds of desires are made possible by new technologies. The creation of new objects of sexual desire and new venues for sexual expression will be facilitated by those who create and exploit new market niches. This will inevitably have consequences for romantic relationships, family formation, and associated close personal relationships.

One can easily imagine that individual hedonic satisfaction could be higher for people who choose to engage with personalized artificially intelligent sex robots than for those who remain directly sexually involved with other human beings. We can also imagine harm at the collective or societal level that would result from the widespread adoption of sex robots. If AI-driven sex robots are widely adopted, the form and function of their agency would be directly relevant to changes in the dynamics of traditional human-to-human sexual relationships. In marriage and dating, changing expectations regarding physical appearance, sexual performance, convenience, etc. are likely to make it more challenging to establish and maintain mutually satisfying human connections based on shared values, vulnerability, and emotional compatibility. One challenge for AI ethics is to articulate the value, if

[5] Those whose moral framework involves an individualistic focus on personal utility might regard social harms as irrelevant or secondary. However, we will assume for the sake of this paper that a radical form of subjectivism with respect to moral matters is either self-undermining or that there are indirect individualist reasons to care about social goods and harms. We are grateful to an anonymous referee for forcing us to be clear on this point.

[6] We are grateful to an anonymous referee for pressing us on this issue and for encouraging us to discuss AI systems that have forms of non-linguistic social agency.

[7] See Gonzalez-Gonzalez et al., 2021 for a systematic review of the scientific literature on sexbots. See also the 2022 special issue of *The Journal of Future of Robot Life* edited by Simon Dube and David Levy on robot sex. Other notable discussions include David Levy's, 2007 book *Love and Sex with Robots.*

any, of the traditional social relations and institutions that are threatened by these technologies.[8]

The use of AI as a substitute for human sexual companionship is likely to contribute to withdrawal from human relationships, especially among those whose bodies, personalities, or subjective sexual preferences are not attractive to others.[9] Sexually intimate relationships typically involve recognition of (and ideally respect for) the other person's interests and desires. However, rather than moderating or adapting subjective sexual preferences in light of the desires and dignity of human lovers, users of sex robots would be free to satisfy themselves without the moral or social demand that they take other persons into account.

From a purely libertarian perspective, this would be a positive development. It would allow users to engage in their preferred forms of sexual expression without harming other human persons.[10] Those who find that their sexual and emotional needs are best met by robots are likely to become less motivated to seek out and invest in the kinds of human connections that emerge from considerate and loving sexual relationships. This would lead to reduced overall social interaction between human persons and would likely have significant secondary social effects.

Embodied social agency of the form exhibited by sex robots is clearly different from the kind of agency that would manifest in a sexual encounter with another human person. The user of a sex robot is subject to different kinds of moral demands and personal risks, and the sex robot itself lacks the kinds of vulnerability and dignity that are present and should be recognized in an intimate sexual encounter with a human being. Clearly, sex robots are an obvious example of the ways in which AI will have morally significant social effects.

3.2 Why Focus on the Linguistically Mediated Social Agency of Chatbots?

As mentioned above, we have found it easiest to focus on chatbots as the core manifestation of the social agency of artifacts while recognizing that linguistic interaction is only part of a complex social story. Nevertheless, linguistic capacity is critically important. For example, in the case of sex robots, the additional agential capacity that is associated with chatbot technology—the ability of artificially intelligent systems to respond conversationally or in adaptive and socially intelligent

[8] Some of these questions are touched upon in Ruiping and Cherry, eds., 2021. Adshade (2017) discusses the economic aspects of social change involving robot sex.

[9] There is a large and growing literature debating the ethics of various kinds of paraphilia and pedophilia as expressed with robots, see for example Jecker (2021), Karaian (2022), Marečková et al. (2022), and Sparrow (2021).

[10] Under these circumstances, our desires to engage in degrading, violent, or simply obnoxious sexual encounters with others or the desire for fully compliant or idealized partners could be acted upon without those desires being brought into question or challenged by the vulnerability and needs of another human person. Of course, for some, the absence of a *real* human person would make it impossible to genuinely satisfy some obnoxious sexual desires given the interpersonal nature of that desire. Sadism, for example, involves the subordination of another human person. It is hard to imagine a sadist enjoying torturing his sex robot for very long, no matter how realistic the robot's expressions of pain might be given the absence of coercion or subordination.

bodily ways—marks the transition from, for example, a mere sex toy or a sex doll to an artificially intelligent sex robot. Linguistic ability is an obvious social matter. For this reason, we focus for the remainder of the paper on chatbots.

Chatbots are software systems designed to serve as conversational interlocutors (Khan and Das, 2018). Typically, this happens through text-based interfaces such as messaging apps, websites, or mobile apps although as we have seen these systems can also be embedded in robots, in virtual or augmented reality avatars, and in other forms of digital companions. Contemporary chatbots engage in sophisticated natural language processing using LLMs to process user inputs and generate appropriate responses. As chatbots become more sophisticated and capable of mimicking human conversation, the degree to which they ought to be regarded as intelligent agents comes into question (Schwitzgebel and Shevlin, 2023).[11] However, as Floridi has noted (2023), it is useful to distinguish agency and intelligence in the ethics of AI. The kind of morally relevant social agency that an AI could engage in would include, for example, conversational manipulation of other social agents, economic exchange and competitive behavior of various kinds, and monitoring and dynamic modification of choice architectures for other social agents. These systems can act autonomously in ways that make a social difference to other agents. As mentioned above, they may also disrupt valuable social relations such as close personal relations. Understanding the nature of this kind of agency is an urgent matter for responsible deployment of AI. In the following section, we examine the range of ways in which the concept of agency has been understood by the engineers who develop these systems.

4 Agency in Computer Science and Engineering

Typically, researchers in AI assume that artificially intelligent artifacts possess agency.[12] In computer science and engineering, the term “agency” figures prominently but with a range of meanings (Jackson and Williams, 2021).[13] At the heart of what we can call the engineering view of agency is the idea that agents are beings that intervene adaptively in their environments. Indeed, in their *Artificial Intelligence: Foundations of Computational Agents* (2017), Mackworth and Poole claim

[11] Traditionally, adult-level human linguistic competence provided a key benchmark for twentieth-century philosophers as they considered the questions of intelligence, agency, and moral standing. Chatbots that run on state-of-the-art LLMs now have the capacity to pass for human interlocutors under certain circumstances, and thus—in the spirit of the Turing test—we are forced to reflect on their level of agency and perhaps even on their moral status. In this paper, we will focus on the question of their agency.

[12] We agree with one of our referees who noted that AI researchers do not simply assume that AI has agency but also presume that the goal of AI is the creation of agents of a certain kind.

[13] According to Wooldridge and Jennings, it was not until the 1980s that the concept of agency received much attention from technologists. They note that “the problem [was] that although the term [was] widely used by many people working closely in related areas, it defied attempts to produce a single universally accepted definition” (Wooldridge & Jennings, 1995, 4). Of course, science fiction has a long history of reflection on the idea of artifacts as agents.

that artificial intelligence should be *identified* with the study of the design of intelligent computational agents. And on their view.

> An agent is something that acts in an environment; it does something. Agents include worms, dogs, thermostats, airplanes, robots, humans, companies, and countries. (2017 4)

Mackworth and Poole's view of AI and agency is typical of the approach to the term "agent" within a broader engineering context. The concept of agent is taken to be intuitively obvious, and it is understood to refer to any entity that can interact with its environment to carry out various tasks. Let's call this the minimal criterion for something to count as an agent. This minimal criterion of agency tends to blur over conceptually difficult questions concerning the nature of action itself. What is it to act? Actions are not the same as mere behaviors; for example, an action is more than a simple reflex or more than merely playing a causal role in some chain of events. Whatever the difference between acting and merely behaving might be, it has typically not been the concern of scientists and engineers.

Talk of artificial agency is never too far from basic philosophical questions, and we can see other computer scientists, for example, Russell and Norvig (2010) beginning to meander into the philosophical swamps with their definition of agent:

> An agent is just something that acts (agent comes from the Latin *agere*, to do). Of course, all computer programs do something, but computer agents are expected to do more: operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue goals. A rational agent is one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome (2010, 4).

In addition to just doing something, Russell and Norvig's conception of agency involves some properties that are philosophically challenging; They follow and create goals, they perceive, they persist, and they are autonomous. In the case of rational agents, they weigh uncertainty and even evaluate alternative outcomes.

These two examples indicate the wide range of features that the engineering notion of agency includes: At one end of the spectrum—the Mackworth and Poole conception of agent—we could count bacteria or plants with relatively minimal agency, while at the other end—Russell and Norvig's rational agent—the paradigm case is the average human adult. Reviewing the literature, uses of the term "agency" in AI research embrace the notion that agency comes in degrees.

Whether it is correct to regard contemporary and near-future AI agency as coming in degrees is a philosophical question rather than a straightforward matter for engineers. As we shall see in the following section, for example, the way that engineers discuss agency contrasts sharply with at least one prominent strain in the philosophical literature of agency. Our view allows for agency coming in degrees but also distinguishes kinds of agency related to distinct kinds of domains for action. In the case of chatbots and companion AI, we argue that they should be treated as agents in the social domain.[14] While chatbots and companion AI might not qualify as agents in some domains and will not have all the capacities that we associate with typical adult human agents, their capacity to act in social contexts is subject to moral evaluation regardless. A key objection to this claim rests on what we call *threshold conceptions of agency*. In the next section, we turn to this important conceptual challenge.

5 The Threshold Conceptions of Agency in the Ethics of AI

Since the 1950s, analytic philosophers have worked to sharpen the concept of agency in part, by determining the necessary and sufficient conditions for distinguishing who or what should count as an agent. Generally, this has involved getting clear on what counts as a genuine action and what distinguishes acting from merely behaving or being a part of some causal chain. The standard view of agency, defended most prominently by Anscombe (1957) and Davidson (1963), held that the very concept of agency relied on some prior notion of intentionality. Some philosophers have also supposed that a necessary condition for genuine agency would include acting *for a reason* or being responsive to reasons in the right kind of ways. Thus, for many who hold the standard view of agency, the concept of agency is inextricably tangled up with desires, mental representations of some kind, and goals in addition to intentions. If something like the standard view were correct, most of the things that computer scientists and engineers think of as agents, including most contemporary AI, would simply not count as genuine agents.[15] Perhaps, the philosopher might argue, this is because AI artifacts are in fact *not* agents, and computer scientists are simply wrong about them. We will explain below why this tempting response impedes careful moral reflection on these systems.

5.1 AI Ethics and Philosophy of Mind

Philosophical discussions of AI agency in recent years have straddled the line between debates concerning the necessary and sufficient conditions for agency and the recognition that AI systems engage with, intervene in the world in ways that appear intelligent, and seem to involve goals, plans, etc. Contemporary AI ethics emerges in part from debates in the philosophy of mind and often concerns itself in detail with questions about whether computers *really* have the kinds of mental properties or states we associate with human minds. Take, for example, Himma's discussion of AI agency. He argues that AI systems could count as agents only if they were first conscious. He writes.

[14] For an informative analysis on how people perceive dogs versus robots as companions.

[15] For a detailed overview of the philosophical literature, see Ferrero, *The Routledge Handbook of Philosophy of Agency*, (2022). For another recent overview on action, see Paul, *Philosophy of Action: A Contemporary Introduction*, (2021).

> it turns out that the capacity for consciousness seems to be presupposed by the simpler notions of agency itself. [...] the concept of agency can be expressed as follows: X is an agent if and only if X can instantiate intentional mental states capable of directly causing a performance; here it is important to remember that intentional states include beliefs, desires, intentions, and volitions (or the relevant neurophysiological correlates. In any event, on the received view, doing A is an action if and only if A is caused by intentional state (and is hence performed by an agent) (Himma, 2009, 27).

Or consider Brey's (2014) and Véliz's (2021) conceptualization of agency and moral agency. Brey and Véliz both argue that artifacts cannot be moral agents because they lack consciousness or higher-order mental states like intentions, beliefs, desires, and the ability to plan. Brey focuses on artifacts in general, whereas Véliz focuses on algorithms. Although Véliz argues that consciousness is a necessary condition for moral agency, she moves between arguing for consciousness as a necessary condition for agency (2021, 593) and consciousness as necessary for moral agency (2021, 490). Brey on the other hand distinguished the two explicitly; he says, "moral agency is a special kind of agency" (2014, 127). Véliz also distinguishes between the two; she says, "tornadoes can also go about their business without human help, and that tells us nothing about whether they are agents—much less moral agents" (2014, 490). Brey on the other hand assumes what we have called a threshold account of agency. He writes, for example, that, "actions are intentional, they depend on capacities for rational thought and self-interested judgments, and the performance of goal-directed behaviors based on such thoughts and judgments" (2014, 127).

We question this formulation of agency. As we and others have argued, agency is best understood as multidimensional and can be studied via a study of its distinguishable sub-capacities. By contrast, Brey, Himma, and others are committed to some version of the threshold model of agency.[16] Though moral agency for LLMs and chatbots is an important topic, in our view, the threshold model blocks the kind of gradated or multidimensional framework that helps shed light on investigations into chatbot agency.

Like Brey and Véliz, we take it as relatively obvious that one can distinguish agency in general from moral agency. However, unlike Véliz and Brey, we reject the threshold account of agency as necessary for understanding agency per se. Here, we do not take a position on whether the threshold account is necessary for moral agency nor do we address the problem of the responsibility gap (Matthias, 2004); both would take us beyond the scope of the current paper.[17]

[16] Silver et al. (2022) also recognize that social agency is best modeled multidimensionally. Although their model primarily tracks level of cooperation between agents, they note, "there are many interactions dimensions critically under researched in relation to Social Agency, and whilst this [their rendition] continuum is centered around the degree of cooperation in an interaction, as Social Agency grows as a field, it is hoped that more key elements will be incorporated into this model" (442).

[17] For an introduction on the issue, see Nyholm (2023), especially chapter 6.

5.2 Why the Threshold Account Makes it Unnecessarily Difficult to Assess Artifact Agency

The views described here all involve versions of what we are calling a strict threshold view of agency.[18] Setting the bar at this level makes most talk of artifact agency implausible as a conceptual matter.[19] Let us consider whether threshold definitions can play a constructive role in inquiry.

We can begin by asking whether versions of the threshold definition of agency are adequate to the task of capturing all cases that we would ordinarily think of as examples of agency. Would dolphins, bonobos, or human infants count as agents? Do neurotypically human adults always meet the threshold for agency? If not, under what circumstances are we genuine agents? Many criticisms of the standard account of agency stemming from Anscombe and Davidson concern themselves with these kinds of liminal cases from animal life or from careful reflection on human cases involving development and disability. Critics have explored cases of these kinds as counterexamples to versions of the threshold model of agency. However, independently of the empirical question of whether some specific definition of agency can make scientific sense of animal, human, or machine agency, we can also evaluate a definition with regard to whether it helps us to adequately think through practical questions in AI ethics. In AI ethics, traditional threshold definitions of agency tend to impede ethical inquiry. To see why, we will present cases that show how threshold definitions make it difficult for the ethicist to both understand and evaluate ways in which computational systems intervene in the world independently of direct human action.

First, consider the ethical assessment of a non-social autonomous artifact. Imagine, for example, trying to determine the level of responsibility that the owner of a Tesla might have for an accident that takes place while the car is in autopilot mode. Most obviously, if we assume that the only agent present during the accident is the human driver, our characterization of responsibilities will differ from the assumption that there are two agents involved. Ethical assessment of the apportionment of blame will depend on understanding whether Tesla was in some sense an agent and whether the software made some decisions during the events leading up to the accident. Notice that its agency need not be conflated with moral agency in the sense that the car itself is an appropriate subject of praise and blame. Instead, the decisions made by the car can be evaluated in relation to the driver's actions. The ability to distinguish the car's decisions from the human driver's decisions is necessary in order to determine the extent to which the driver should be blamed for the accident and the extent to which the developers of the software are morally responsible.

[18] For an overview of the logic of threshold arguments in the study of cognition, see Calvo and Symons (2014).

[19] Of course, those who hold the threshold account might retreat to some kind of instrumentalist conception of artifact agency. We can certainly act as though an artifact is an agent for instrumental reasons in the spirit of Dennett's intentional stance (see Symons (2001) and Dennett (1987)), but given this version of the threshold view, we cannot ascribe agency to artifacts like chatbots independently of an observer's ascription of agency. We will return to this option below.

Determining how to assign responsibility in the case of the self-driving car will clearly depend on multiple factors related to its level of agency.

Nylhom (2018) makes a similar point. He notes that different types of agency are relevant when assigning responsibilities. The domain and the context in which the AI system is being used partially determine the type of agency or level of responsibility involved. The humans in question can be the driver or the programmer/engineer. Nyholm concludes we should be careful not to overattribute agency to AI systems, and it is far better to think of AI agency "occurring within human–robot collaboration" (2018, 1218).[20] We fully agree. However, in order to properly characterize that collaboration, we need to characterize the kind of agency that the artifact possesses.

The threshold model of agency makes it difficult to assess the ways in which artifacts agency is multifaceted and context-dependent. An AI system that acts in social systems in agent-like ways might be utterly without agency in other contexts. Or take, for example, the software system in a self-driving car. It may possess different degrees of agency depending on the context. For example, the car may possess high levels of spatio-temporal agency in terms of its ability to make decisions and perform tasks that are beyond the ability of a human being.

A car's anti-lock braking system (ABS), for example, is a safety feature in modern vehicles that prevents the wheels from locking up during hard braking, allowing the driver to maintain steering control and avoid skidding or sliding. The ABS control module receives information from the wheel speed sensors and determines when a wheel is about to lock up. When this happens, it sends a signal to the hydraulic modulator. The hydraulic modulator controls the brake pressure applied to each wheel. When the ABS control module sends a signal, the hydraulic modulator rapidly applies and releases the brake pressure to the affected wheels to prevent them from locking up and going into a skid. The ABS exhibits a capacity that outstrips our own unaided human agency. This is because the human agency does not extend to the kinds of micro-level adjustments to brakes that happen in fractions of time below the threshold of human consciousness. The agency of the driver who applies the brakes can be understood to be extended or supplemented by the actions of the ABS. However, the driver cannot be not held morally responsible for decisions that happen in the control module of the ABS system that take place in a time scale that falls below the threshold of human consciousness.[21] This example indicates that assignments of moral responsibility cannot be straightforwardly wedded to the kinds of agency we associate with normal human psychology and abilities.

[20] One issue with this is that as highlighted by Silver et al. (2022, 449), several psychological studies have demonstrated that joint action or joint agency is difficult to justify between robots and humans. Humans tend to not think or report a sense of joint agency when collaborating with robots. Also see Nylhom (2023): Nylhom spends nearly an entire chapter (chapter 3) in his book (2023) on the various moral issues and approaches for autonomous vehicle. Also, see chapter 4 of the same book for further debates on autonomous cars.

[21] Floridi and Sanders also make a point to underscore the difficulty of holding humans responsible for computing systems (AI, regular software, and so on), features, or actions unforeseeable by humans (2004, 371–372) (CITE), like our example of the ABS system in cars.

Similarly, the level of agency of a self-driving car may vary depending on the specific context and situation. For example, a self-driving car may have a high level of agency on a highway where there are few variables to consider but a lower level of agency in a busy city center where there are more obstacles, traffic, and unpredictable human behavior. Changing contexts will impair the capacities of the autopilot system to make reliable and competent decisions. While the same considerations apply to human drivers, we must be able to understand degrees of agency to properly assess the appropriate allocation of praise and blame in these contexts.

In addition to recognizing degrees of agency, understanding how to assign responsibility also involves being able to distinguish kinds of agency. So, for example in the Tesla, the car's autopilot system will exhibit low (perhaps no) *social* agency in terms of its ability to interact with other drivers, pedestrians, and police. So far, for example, it has not seemed able to readily adapt to varying cultural norms related to traffic and decision-making in different societies. Norms around how one should drive in the presence of emergency vehicles for example are highly context sensitive and involve common knowledge considerations that are likely going to prove challenging for AI systems. For example, if a car is on a narrow and busy street in a city like Lisbon or San Francisco, other drivers and pedestrians will act on common knowledge of what one does when an emergency vehicle needs to pass. Pedestrians in Lisbon might expect a car to come very close to a crowded sidewalk or to even partially drive onto the sidewalk in order to allow the emergency vehicle to pass. In San Francisco, the norms will be somewhat different, and pedestrians might be alarmed by a car coming close to them on a sidewalk.

Sensitivity to the contextual nature of agency is critically important in assessing assignments of responsibility in cases like this. Understanding this helps us to assess the developer's role in the design of these systems. Whatever agency exhibited by self-driving cars obviously depends in part (but not entirely) on the design and programming choices made by its developers. For example, a self-driving car may be programmed to prioritize safety over speed, which may limit its decision-making capacity in certain situations. Another frequently discussed trade-off in the development of self-driving cars is the need to make decisions that prioritize the safety and well-being of passengers versus the safety and well-being of other people. For example, if a pedestrian suddenly walks into the path of the car, the car may need to decide whether to swerve and potentially endanger the passengers or hit the pedestrian. Advocates of the threshold view of agency would certainly not wish to completely block assessments of moral culpability in these cases, but as we have seen, using human psychological criteria for agency makes actual cases, like AI in autonomous vehicles, more difficult to assess. If we admit degrees and kinds of agency, we will be able to identify morally relevant decision points and explain why a system opted for one course of action over another. This is a necessary condition for reliable assessments of responsibility.

5.3 Taking Kinds and Degrees of Agency into Account in Moral Assessment

Previous philosophical discussions of the ethics of autonomous vehicles frequently involve questions related to the agency of these systems. Woollard, for example, reimagines the trolley problem with self-driving cars. She examines the difference between doing and allowing harm and whether autonomous vehicles can be considered moral agents. In a typical trolley case, the ethical question revolves around doing or allowing harm, but the agent is a human. For Woollard, how we think about full self-driving or automated-hybrid driving in trolley-problem scenarios depends on "(a) our conception of the behavior of driverless cars [their agential status]; (b) the forms of driverless cars that are developed and used; (c) the background expectation of programmers/manufacturer/owners of driverless cars and the conditions of being able to put those cars on the roads" (51). We understand Woollard as endorsing the idea that agency should be understood as coming in degrees or on a spectrum (see also Nylhom and Talbot et al. (55–56)).[22] Woollard's focus is specifically on moral agency, and she evaluates different views of the agential status of autonomous vehicles in relation to the ways that we would typically evaluate harms that result from their use. While we are not focused on the moral agency of chatbots in this paper, Woollard's approach to the question of agency is applicable to social contexts.[23]

Mecacci et al. (2023) develop a pluralistic moral responsibility framework to challenge the dichotomy of automated vehicle control vs. human agent control. They call it meaningful human control (MHC) (2023, 1156). They argue that we ought to spread responsibility across actors, going beyond the agents (artificial and human) directly involved. Under their framework, the designers and policymakers all the way down to pedestrians and cyclists navigating the streets end up sharing responsibility for harmful events involving autonomous vehicles. The ways that agents behave and the events that take place in urban traffic scenarios are combined into what they call a "system." The "reasons that move the system (urban traffic), both moral and practical ones, must be clearly identifiable, together with their human carrier(s)" (1161). The rules governing whose (policy maker, pedestrian, cyclist, driver) reason the system should respond to at different points in time are thought of in a flexible and situation-dependent way. They conclude that, "the pyramid [hierarchy of agents – cyclist, driver etc.] is meant to provide the necessary categories for a behavioral rule that dynamically maximizes MHC across all the users, thereby providing with more circumstantial – and more intelligent – solutions to coordination problems" (1162).

While previous discussions of autonomous vehicles have pointed to new ways in which the agency of artifacts can be understood, they are generally still attached to a threshold model of agency. This is understandable in discussions concentrating on questions of moral responsibility and moral agency. An agent either is or is not morally responsible for its action. That responsibility might be shared or distributed, but the key question (correctly) in these discussions is whether an artifact meets the

[22] Also, see chapter 2 of Nyholm (2020).

[23] We thank an anonymous referee for encouraging us to distinguish between moral agency and agency per se.

threshold for counting as morally responsible. Thus, it is unsurprising that in the context of autonomous vehicles, philosophers have typically treated moral agency in ways that assume a threshold model of agency. However, returning to our Tesla case, above, when we approach the assignment of ethical responsibility, we must be in a position to characterize varying degrees of agency and decision-making capacity in the vehicle in order to know whether the driver, the developer, or someone else is culpable.

As we have argued above, properly evaluating social agency in AI will require us to distinguish the kinds and degrees of agency. When we considered the cases of the ABS system, we noted the importance of distinguishing the agency of artifacts from human psychological capacities. We also noted the importance of dimensions and degrees of agency in the discussion of the capacities of an autonomous vehicle on a Mid-Western Interstate highway in the USA when compared with an autonomous vehicle driving in Bucharest or Boston. These examples were intended to highlight the moral significance of carving agency up in contextually sensitive ways.

When we think of social agency, we will find a similar range of different kinds of autonomous AI actions—from purely linguistic, conversational interactions to socially relevant actions that take place at timescales faster than human beings can detect, to social interventions that involve amounts of data and computational processing power that exceed our abilities, and to population-level interventions that are undetectable at the individual level. The effects of AI social agency will be felt at a range of different scales and levels of abstraction. An AI might have a range of social consequences for individuals, but also for the continuity and efficacy of certain social norms, the resilience of certain social institutions, the well-being of populations, and the like. There will be a diverse range of autonomous agents in social systems, and it is imperative that we open our theoretical approach to agency accordingly if we hope to effectively evaluate the social character of AI.

Following the example of Calvo and colleagues who describe the emergence of what they call *minimally cognitive agents*, we will introduce an approach to agency that shows how we can unpack distinct components of artifact agency. This will allow us to identify ways in which an artifact can, for example, act in adaptive ways without being conscious and without having its own intentions, mental representations, etc. While we believe that such an approach will have benefits for considerations of *moral* agency of the kind Mecacci et al. and Woollard are undertaking, defending that position is beyond the scope of this paper.

6 The Emergence of Minimal Social Agency: How Far Can We Lower the Bar?

What is required for an account of social agency that would shed light on AI's interventions? Assuming that most threshold accounts simply deny that AI artifacts really are agents, the first option is to relax our requirements for what can count as an agent. Lowering the bar can only go so far of course, and there are a range of competing conceptions of agency in which philosophers have emphasized a set of minimal conditions for the explanation of agential and cognitive capacities in, for example, simple biological systems or in the development of these capacities in children (Burge, 2009; Calvo et al., 2014; van Hateran, 2015, 2016; di Palo, 2005).

6.1 Core Conditions for Minimal Agency

Barandiaran et al. (2009) identify three core conditions for a minimal agency. Firstly, the system must possess some level of individuality—perhaps maintaining enduring boundaries that distinguish it from its environment (Barandarian, 2009). Secondly, it must be a source of activity within its environment that is characterized by an interactional asymmetry. And thirdly, it must be capable of regulating its activity in relation to certain norms or rules. Even basic forms of proto-cellular systems can fulfill these conditions and would count as examples of minimal agency in this view. In the case of AI, we can find some or all of these minimal behavioral characteristics present in artifacts in some contexts.

Before examining the way these characteristics figure in chatbot social agency, it is worth spelling out the role of interactional asymmetry in the minimal concept of agency: While a cloud is not an agent, a bird gliding (with minor movements) through the air is. The motion of the cloud depends entirely on its internal physical and chemical properties and the physical environmental laws and properties governing those relations. Furthermore, the cloud does not distinguish itself from its environment. By contrast, the bird is highly sensitive to the distinction between itself and the environment and exerts forces in order to modify its flight path.[24] It can adaptively modulate its relationship with the external environment as it glides, and its flightpath is not entirely dependent on forces that are independent of it. Tyler Burge (2009) argues for a similar interactional asymmetry condition. Like Barandiarian et al., Burge also requires that the whole organism carry out the behavior as a unit and not by subsystems (Burge, 2009, 261, Barandarian et al., 281). This requirement might not be applicable in the case of an AI artifact.

The modulating relationship with the environment is at the heart of the second condition. The entity must be "capable of engaging in some modulations of the coupling and doing so at certain times but not necessarily always" (2009, 372). The term *coupling* in this case refers to the bird's interaction with its environment. In the case of a cloud, the interaction between the cloud and its environment is symmetrical because the cloud "cannot constrain [at least some of the conditions] this coupling in a way that the environment (typically) does not" (Barandarian et al., 2009). The entity must be able to break the symmetry of its coupling (relation with the environment) from within, thus satisfying interactional asymmetry conditions.

Minimal agency is not restricted to living organisms, since for Barandiaran et al., agents are simply identifiable systems that adaptively regulate their coupling with their environments in a regular or rule-guided manner. While a range of other conceptions of minimal agency have been defended, they all roughly converge on

[24] Consider what van Hateren says concerning conditions required for minimal agency, "such conditions should indicate which species have agency and which behaviors are *acts* [emphasize ours] rather than something else (… such as sneezing, shivering [automatic reflexes]".

something like these three properties, and all offer an alternative to the view that it is necessary for an entity to possess full, adult-level, neurotypical human cognitive abilities to be considered an agent. Minimal accounts of agency can avoid the restrictiveness of the threshold accounts because they are able to capture its adaptive and asymmetrical features while leaving room for variation and complexity in actual human and non-human actions. Rather than taking markers drawn from typical adult human psychology as definitive of agency, a minimal approach focuses on the systematic properties of agents while admitting a variety of ways in which those systematic properties can be realized.[25]

As we have seen, if an account of agency is too demanding, it will exclude certain types of behavior or individuals whom we recognize as agents. For example, if an account of agency requires that an agent always acts with a certain degree of rationality or autonomy, it may exclude individuals with certain cognitive or neurological disabilities, who may not meet those criteria but are still capable of intentional action. Similarly, intentional action can occur without rationality. Intentional action refers to an action that is performed with a specific intention or goal in mind. It involves the agent's ability to direct their behavior toward a particular end, even if that agent is not necessarily rational or logical.

A young child or a chimp may intentionally grab a toy from someone else because they want to play with it, even though they may not have the capacity to fully understand the consequences of their actions or the rationality to appreciate the implications of taking the toy. Similarly, individuals with certain neurological or cognitive disabilities may engage in intentional actions even if their behavior is not always rational in the traditional sense of the term. They may still possess the capacity to form intentions and act on them, even if they lack the ability to fully comprehend the rational implications of their actions.

Minimal accounts of agency can be comprehensive enough to capture the basics of agency in ways that allow for a scientific explanation of the emergence of agential capacity in its more sophisticated forms. Here, we are drawing on arguments in recent philosophy of cognitive science where some philosophers have worked to characterize different aspects of cognition as it appears in non-human organisms (Calvo et al., 2014). Calvo and colleagues, for example, see plant cognition as exhibiting some aspects of cognition while being far removed from normal adult human cognition. Such systems can be understood to be *minimally cognitive* in their view. For Calvo et al., it is obviously not the case that minimally cognitive systems have human adult–level beliefs and desires. Instead, certain capacities and properties of, for example, the behavior of a plant or a nematode contribute some (but obviously not all) aspects of what we identify

[25] Debates around group agency are also worth noting here. Groups per se lack any representational content or reflective thought but do seem to take actions which, at least, seem irreducible to individual members (parliament voted to do X). For informative and contrasting view on group agency, see Lewis-Martin (2022). It is worth noting that some philosophers (Christian List, 2021) have characterized AI agency as similar to group agency—List argues that AIs are agents by drawing parallels with group agency. Group agency is a contentious topic, and nothing in our current argument rests on accepting it. We mention it here to note the possibility of agency without intentionality or at least without intentionality in the conventional sense.

with human adult–level cognition. We follow the approach taken by Calvo et al. to the emergence of minimally cognitive systems in the context of agency. By analogy, it is vital, in our view, to recognize that different degrees and dimensions of agency in artifacts and simple biological systems admit consideration and have their own distinctive sets of normative implications.

Some versions of the position we defend here with respect to minimal agency can be found in previous works (see Ferraro (2022), Nyholm (2018)). As mentioned above, Nyholm is concerned with assigning moral responsibility. Similarly, Strasser considers degrees or dimensions of agency as a notion relative to the moral agential status of AI systems. Strasser evaluates the distribution of moral responsibility between human and AI systems in joint action scenarios. Like us, she also highlights the difference between gradated and conventional philosophical accounts of agency.

An AI system must be a social agent for it to be a moral agent, but social agency is a necessary but not a sufficient condition here. Strasser points to animals as exemplifying this type of decoupling of social and moral agency (2022, 526). For Strasser, an AI system is a social agent "if artificial agents contribute to social interactions by utilizing socio-cognitive abilities and thereby add to a reciprocal exchange of social information, we are justified to consider them social interactions partners" (Strasser, 2022, 524). This in turn for Strasser can justify minimal moral agency for AI systems under her distributive framework for moral responsibility (ibid, 2022, 527). Similarly, Ferraro writes, for example, "it might be that some dimensions of agency could be attributed only to more complex organisms but not to simpler ones. If so, what are these dimensions? How are they related to each other? What are the normative implications of these attributions?" (Ferraro, 2022, 6). In our view, Ferraro's questions are precisely those that need to be answered when we consider the social agency of AI, and they are questions that would be preempted by the traditional threshold account of agency.

At this point, it should be clear that minimal or multidimensional accounts of agency will serve very different philosophical purposes than the threshold account that emerged in analytic philosophy from the 1950s onward. Rather than being focused on definitions that carve out a familiar domain of neurotypical human adult–level experience, the purpose of the approaches and accounts we favor is typically to contribute to some explanatory project related to either the emergence of different forms of agency or as we saw in the discussion of autonomous vehicles to illuminate some other domain of interest, either technological development or the apportionment of moral and legal responsibility.

7 What is Required for a Chatbot to Have (At Least) Minimal Social Agency?

What would it mean to characterize chatbots as minimal agents in the sense discussed above? At this point, we have argued that they can exhibit social agency despite not possessing important features of neurotypical adult human agency. But what does it mean to say that chatbots perform social actions? Our contention is that chatbots intervene in society via their interactions with human persons and with relevant kinds of social institutions. Chatbots intervene in existing social relations in a range of ways. Economic transactions, sexual relationships, friendships, and political life are already being affected by the influence of AI-powered conversational systems.[26] In this sense, we can plausibly regard chatbots as acting via conversation in the social world.

How and whether an agent can exert control partially depends on the nature of its environment. The context within which chatbots are embedded is social and conversational. Chatbots have influence over others and differentiate themselves from other entities at the level of language and social relations.[27] The fact that much of what chatbots do is conversational means that their effects on their environment are partly governed by the differences between themselves, the chatbot, the immediate user, other users, non-users, and so on. If a conversation is happening, then there are distinguishable interlocutors involved. One can deny that a conversation is really taking place in any particular instance of human interaction with a chatbot, but that seems like an implausibly strong position to take. If one is convinced, for example, that chatbots are not agents and that in any genuine conversation, one's interlocutor must be an agent, then one might be driven to the conclusion that we never actually chat with chatbots. However, such an exclusion is question-begging. It is better to assume a common sense position that there often is a conversation of some sort taking place and that there is some prima facie case for saying that the chatbot can count, at least minimally, as a being with a distinguishable identity.[28] Recognizing the conditions governing conversations at the social level allows us to see how chatbots might count as at least minimal individuals.

[26] For example, when a user engaged with a therapy agent in a conversation, for the human, even if they know the interlocuter is an AI, the perception of the outputs of the chatbots for the user is perceived as conversational actions. Take the example by Yang (2020); the user says to a chatbot "Hey, I know you are not real, but I just wanted to send these pictures of my family out at Disneyland having a great time. I'm doing better now. Thank you" (35). The user seems to take the chatbot as an agent worthy of respect that they should be polite and share intimate family details with. Another example is the language use around ChatGPT or midjourney. It is common to see headlines or conversations that have language like, "what does chatGPT think X is, or this is what AI thinks people from Y country look like." A person in Korea legally married a virtual avatar (Jozuka et al., 2018). Robotic animals, like Paro, have been around for a while, or for our case, the ChatGPT induced pet bots like Loona. One final example to demonstrate the inclusion of AI systems like chatbots. Take the prevalence of friendbots like Replika. During the pandemic, reports of using chatbots like Replika for therapeutic reasons (Weber-Guskar, 2022) were up. As mentioned, there is a growing acceptance of using chatbots or LLM-equipped robots as sexbots.

[27] One of our referees noted that it might be helpful to think of *the social* by reference to Floridi's concept of levels of abstraction (LoA) (2006, Floridi & Sanders, 2004). By using abstraction, one can further clarify a particular phenomenon or artifact of inquiry by focusing on one set of properties or detail over another set. Usually, one set is more abstract than the other. This permits researchers to focus on a particular aspect of the inquiry for different purposes or to be more explicit about the goals of particular explanations. Floridi puts it as follows: consider the wine example. Different LoAs may be appropriate for different purposes. To evaluate a wine, the "tasting LoA," consisting of observables like those mentioned in the previous section, would be relevant. For the purpose of ordering wine, a "purchasing LoA" (containing observables like *maker*, *region*, *vintage*, *supplier*, *quantity*,and *price*) would be appropriate, but here, the "tasting LoA" would be irrelevant. In our case, we can focus on the social LoA: the level of conversations between two entities and the socio-linguistic world.

[28] Here, the conditions governing the individuality (rather than the identification) of the artifact come into play. Here, see Symons (2010) for a discussion of the individuality of artifacts and organisms.

But what about interactional asymmetry? Chatbots operate within the social environment and negotiate their responses given embedded constraints in their programming. For example, take Sparrow (chatbot), which has 23 built-in rules for ethical conversations, or ChatGPT, Bard, and BingChat. They all have some built-in normative rules by which they abide. Chatbots modulate linguistic responses and shape conversations based on their internal constraints. In this sense, like biological agents, chatbot action exhibits some degree of interactional asymmetry. Chatbots modulate their responses while adhering to the expectations that are shaped by the rules and norms of the social environment. Thus, they satisfy the second and third conditions provided by Barandiaran et al. (2009).[29]

Chatbots exhibit this minimal form of agency at the social or conversational level. Clearly, as we have noted repeatedly, there are important differences between the agency of an adult human being and a chatbot; this applies also to social agency. Chatbots generally do not occupy locations in existing social hierarchies for example. They do not have the kinds of varying social status that we would ascribe to other human beings. Human social relations involve status, power, subordination, and other kinds of hierarchical phenomena. While chatbots can pretend to occupy locations in the social hierarchy (they can serve as assistants or servants), the meaning of their fictional location in human hierarchies is not equivalent to human cases. This is, of course, a highly complex topic and deserves further study.

7.1 But are Artifacts Really Agents? Beyond Common Sense Threshold Questions

Human social action is not limited to conversation alone. As discussed above, carebots and sex robots already interact physically in ways that are unambiguously social, albeit with limited capacities and sophistication (Friston et al., 2021). It is also the case that recent work on generative agents (Park et al., 2023) has simulated distinct interacting agents in a game-like environment. Park et al. have created an environment in which chatbots are virtually embodied in a Sims-like game that includes simulated geographical location, separation between agents, and spatial encounters between distinct agents. Thus, while we focus here on the social agency of chatbots, they have the potential to operate in other domains and possess different kinds of agency.

It will soon be natural to ask, from a common sense perspective, whether we ought to regard these systems as *genuine agents*. This common sense threshold

[29] Like Barandiarian et al.'s, conditions for minimal agency, Floridi and Sanders (2004) also provide base conditions for agency—(a) interactivity, responds to environmental stimuli; (b) autonomy, governs its behavior independent of environmental stimuli; and (c) adaptability, modify its past system states and transition rules according to the environment taking into account success and failure of task (357- 358, 363–364). These conditions are similar Barandiarian et al.'s. Autonomy is similar to the individuality, adaptability and interactivity have parallels with interactional asymmetry, and adaptability is akin to normativity (success or failure at achieving normative goals). Of course, these conditions are not exact replicates. Also, like Floridi and Sanders, we highlight the importance of LoA for chatbot agency. Chatbots are best understood as agents when viewed at the social or linguistic LoA. Although Floridi and Sanders differentiate between agency and moral agency, their ultimate goal is to establish moral agency for AI systems by first showing their agential status.

demand is natural, but as we have seen, it puts several subtle questions that are raised by the idea of artificial social agency off limits. Some of these have been discussed in depth by the human–robot interaction community. Jackson et al. (2021, 7–8) for example distinguish perceived social agency and actual social agency. They argue that perceived agency from the user's point of view is relatively uninformative with respect to the system itself and that any actual agency is better analyzed from the developer's point of view.

Ultimately, judgments with respect to the agential status of chatbots cannot be entirely a matter of the opinions of users. Jackson and Williams are certainly correct to say that there are at least two levels at which chatbot agency can be understood. However, it strikes us as too quick to contrast the naïve user-level perspective (which we discuss below) and the sophisticated developer level (the realistic level). Developers may be surprised by the manner in which and the degree to which their creations play a role in the social world. Nevertheless, we agree that we ought to be sensitive to the overattribution of agency.

Returning to our ordinary experience of chatbot social agency, chatbots can demonstrate high levels of linguistic competence and reasoning ability but sometimes fail to demonstrate basic kinds of understanding in ways that—until relatively recently—have made it obvious that we are not engaged with a conversation with a human person. At present, these systems can engage in complex reasoning but sometimes fail to recognize aspects of problems that we might find obvious or simple. They can respond impressively and in flexible ways while at the same time being unable to modify their goals or to rethink the purposes of their actions. They seem adept at some tasks that we associate with agency while failing in other dimensions. However, as they develop, obvious failures will become less frequent, and they will likely outstrip our intuitive capacity to recognize them as non-human, at least under certain conditions. For example, GPT 4.0 has been put through various cognitive tasks to test its reasoning capacities. In a recent study, Ullman (2023) tested GPT 4.0's on theory of mind (ToM) tasks. He found that GPT 4.0 fails a trivially altered version of ToM tasks. On the other hand, one can have lengthy and in-depth conversations with GPT 4.0 or even ChatGPT, and remarkably, the system is able to keep the conversation going.

Pre-theoretical experience of people using chatbots is a difficult and complicated empirical matter. Users recognize the deficiencies of the LLM and yet slip easily into thinking of them as more than mere software systems (Brandtzaeg et al., 2022; Gillath et al., 2023). Given the ability to mimic the kinds of responses that we expect from adult humans, it is natural that we perceive these systems as intelligent and as possessing agency. As Dennett argued, taking an intentional stance in relation to such patterns of experience is almost unavoidable (1987). It is entirely predictable that human beings will tend to attribute intelligence and autonomy to well-designed chatbots, independently of whether we would assent to the explicit claim that these systems can really reason, act intentionally, possess a complex inner life, or make decisions (See AU 20??). Likewise, we may also ascribe emotions and subjective experiences to chatbots, perceiving them as having their own purposes and desires.[30] Indeed, we even tend to regard chatbots as moral agents of some form such that when a chatbot makes an error or fails to perform as expected, we may be inclined to blame the system for its actions, as we would a human person. Again, these tendencies can be *felt* independently of whether we *genuinely believe* that chatbots have the relevant features that might make it reasonable to praise and blame chatbots for their actions.

Our tendency to ascribe full adult human agency to chatbots is heightened by the practice of creating chatbots that refer to themselves with first-person pronouns, that mention histories that they do not have, that pretend they have bodies, etc.[31] Most developers build chatbots with the goal of engaging users and causing a feeling of trust. What this means in practice is that developers are leveraging our natural tendency to ascribe intentionality in order to manipulate people into believing that they are interacting with something like a human being.[32]

Sedlakova and Trachsel (2022) offer an approach to understanding chatbots' agential status in these settings—a hybrid between a tool and an agent (2022, 6). Sedlakova and Trachsel (2022) also endorse a threshold account of agency or what they call subjects (ibid, 7–8). However, like Véliz (2021) Sedlakova and Trachsel do not differentiate between moral agency and agency per se (footnote 2, 2022, 6. Blinkley and Pilkington (2023)).[33] For an account of moral agency for any AI system like chatbots, clarifying their agential status is required. Moreover, contrary to Holohan et al. (and HRI developers), the agential status of chatbots and AI systems cannot depend on the perception of humans or their relationship with humans.

Inevitably, approaching the agency of artifacts in the way we recommend will put us at odds with people's common-sense reactions when interacting with chatbots. The human tendency to over-attribute agency to the world will make it difficult for

[30] Shanahan (2023) underscores this point for LLMs, as he says, "a bare bone LLM [for instance] doesn't really know anything because all it does, at a fundamental level, is a sequence prediction" (2023, 5). So, although it is tempting to ascribe intentionality, beliefs, and desires to these systems, it is a mistake. For Dennett, the intentional stance was understood to be an adaptive trait to specific environmental and evolutionary pressures. In this sense, we are "right" to ascribe beliefs and intentions to aspects of the world that evolution shaped us to detect. See AUTHOR 20?? for a discussion of the relationship between the appropriateness of taking the intentional stance and Dennett's skepticism with respect to realism about representations and intentions.

[31] Not all chatbots are deliberately deceptive in this respect. In 2022, Sparrow AI from Deepmind was explicitly built to avoid this kind of deceptive action in relation to users. Their working paper provides a detailed description of the heuristics that they employed to guide their chatbot (The Sparrow Team, 2022).

[32] Nyholm shares our criticism of demanding accounts of agency and seems to endorse some version of the view we defend here (2018, 2023).

[33] Similarly, van Lingen et al. (2023) also affirm threshold approaches and slip between moral agency and agency simplciter. For example, Blinkley and Pilkington say that to be a minimal agent is "to simply [perform an] intentional action" (2023, 25), and van Lingen differentiates between strong and weak AI (2023). For van Lingen, chatbots are weak AI. Strong AI can have phenomenal experiences, but weak AI cannot; therefore, weak AI is not a moral agent (22). Furthermore, weak AI, for Lingen, cannot act without human actors; thus, they cannot be agents (23). Some, like Huber, take a different approach. Huber suggests that the pragmatic benefit of AI is more important than whether they are actual agents or not. Lastly, Holohan et al. (2023) suggest that agency in therapeutic contexts emerges as a result of the relationship between chatbots and patient (15).

us to resist experiencing chatbots as beings whose agency is on par with normal human adults. Chatbots possess linguistic capacities that allow them to engage in social activities like conversations. In that sense, they can exhibit social agency of some kind. While almost everyone agrees that linguistic capacity of the kind exhibited by contemporary LLM-driven chatbots affords an entity to a higher degree of agency than non-linguistic ones, the matter is complicated in the case of artifacts.[34]

They can provide us with post hoc linguistic discussions of reasons for their actions, but arguably, they lack the kind of internal architecture to reflect on those reasons in the ways that an adult human might. It will be important to distinguish ways in which chatbots exhibit agency while avoiding either mistakenly elevating their agency or adopting a prematurely dismissive attitude on a priori grounds. However, by considering the idea of minimal agency as discussed above, we can find additional options, focusing instead on the component features of minimal agency as exhibited by some AI systems as described above and recognizing those components as having morally relevant features.

8 Conclusion

Agency, we have argued, comes in kinds and degrees. We have emphasized the social agency of artifacts as a key topic for AI ethics and have offered a range of reasons both for taking the issue seriously and for thinking about the component aspects of agential behavior as having morally relevant features. Rather than dismissing agency that does not meet the requirements of what we call the threshold accounts, we suggest that there are productive ways to think about more minimal forms of agency than full adult human agency. We saw that agency comes in kinds and that some entities like animals are non-linguistic but social agents (Glock, 2009, 2019; Steward, 2009). Differences in kinds of agency are related to varying cognitive capacities but also to varying ecological niches and contexts. Turning to AI, chatbots clearly lack some of the cognitive capacities of animals and are neither embodied nor sentient. However, they do possess high levels of linguistic competence. As we have seen, their interventions in the world not only affect individual users but are also disruptive to social conditions. Chatbot technology has served as our focus here because it is the most obvious *social* aspect of social AI. As noted in Sect. 2, we recognize that there are aspects of social relations that go beyond the straightforward linguistic. As a result, our comments on the social agency of AI are incomplete, and it is important to acknowledge that there are complex emotional aspects to the social role of AI that fall beyond the scope of this paper. We recognize with Loh and Loh that as robots move from "dull, dirty, and dangerous" industrial and military applications to more social domains, their roles "are marked by increasing levels of interdependence, and physical and emotional closeness".

[34] Glock provides an overview of the reasons philosophers deny that animals act. The primary basis is the claim that animals do not act in virtue of reasons (Glock, 2019, 667).

AI enthusiasts and developers often complain that philosophers move the goalposts or change the standards according to which these systems should be judged. This is a reasonable complaint, and part of our message in this paper is that it is unhelpful for AI ethics to insist that AI has failed to meet some elusive standard for *genuine* agency. Rather, we ought to recognize that the concept of agency has many distinct parts and that many instances of AI should already be regarded as agents. This is especially true in the domain of social agency. We repeatedly noted that recognizing the agency of chatbots does not compel us to regard them as having the same capacities or moral standing as human beings. For example, it does not necessarily mean artifacts themselves should be subjects to praise or blame; they can be agents without being moral agents. As we have seen, the distinction between agency and moral agency is frequently ignored in the AI ethics literature. Instead, we have argued that it is more illuminating to distinguish various aspects of agency. Our analysis provides a framework wherein we can distinguish the different aspects or properties of agential activity. We adopt an approach analogous to that taken to minimally cognitive agency in the study of plants and simple animals. That approach to cognition served as a model for the methodological strategy of isolating specific features of adaptive, autonomous, and rule-guided behavior that artifacts can manifest.

Without a clear understanding of the kind and degree of agency that chatbots and other social AI possess, it will be difficult to determine their legal and moral status and the responsibilities of their owners and developers for the reasons we discussed above. Tackling conceptual questions concerning the degrees and kind of agency for artifacts will help guide the research, development, and governance of autonomous technologies. While we have focused here on chatbot agency and have emphasized social agency, our approach is applicable to a broader range of artificial intelligence applications in different domains.

Acknowledgements We gratefully acknowledge the excellent feedback from two anonymous referees. Conversations with the AI ethics graduate seminar at the University of Kansas in 2021 formed the basis for this project in addition to helpful discussions with Ramon Alvarado, Oluwaseun Sanwoolu, Oluwakorede Ajibona, Francisco Pipa, Luciano Floridi, Jack Horner, Amir Modarresi, John Sullins, and Caroline Arruda.

Authors Contribution The authors equally contributed to this study.

Funding Ripple,U.S. Department of Defense, H98230-23-C-0277, John Symons.

Data Availability We do not analyze or generate any datasets, because our work proceeds within a theoretical and mathematical approach.

Declarations

Competing Interests The authors declare no competing interests.

References

Adshade, M. (2017). “Sexbot-induced social change: An economic perspective.” In *Robot Sex: Social and Ethical Implications*, 289–300. MIT Press.

Anscombe, G. E. M. (1957). *Intention*. Basil Blackwell.

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior, 17*(5), 367–86. https://doi.org/10.1177/1059712309343819

Binkley, C. E., & Pilkington, B. (2023). The actionless agent: An account of human-CAI relationships. *The American Journal of Bioethics, 23*(5), 25–27. https://doi.org/10.1080/15265161.2023.2191035

Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research, 48*(3), 404–429.

Brey, P. (2014). From moral agents to moral factors: The structural ethics approach. In P. Kroes & P.-P. Verbeek (Eds.), *The Moral Status of Technical Artefacts* 17:125–42. Philosophy of Engineering and Technology. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-7914-3_8

Burge, T. (2009). Primitive agency and natural norms. *Philosophy and Phenomenological Research, 79*(2), 251–278.

Calvo, P., & Symons, J. (Eds.). (2014). *The architecture of cognition: Rethinking Fodor and Pylyshyn's systematicity challenge*. MIT Press.

Calvo, P., Martín, E., & Symons, J. (2014). The emergence of systematicity in minimally cognitive agents. *The architecture of cognition: Rethinking Fodor and Pylyshyn's systematicity challenge*, *397*.

Davidson, D. (1980). *1963, "Actions, reasons, and causes", reprinted in Davidson Essays on actions and events* (pp. 3–20). Clarendon Press.

De Gennaro, M., Krumhuber, E. G., & Lucas, G. (2020). Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in psychology, 10*: 3061.

Dennett, D. C. (1987). *The intentional stance*. MIT press.

di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences, 4*(4), 429–452. https://doi.org/10.1007/s11097-005-9002-y

Ferrero, L. (Ed.). (2022). "Introduction." In *The Routledge Handbook of Philosophy of Agency*, 1–18. Routledge Handbooks in Philosophy. Abingdon, Oxon ; New York, NY: Routledge.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4358789

Friston, K., Moran, R. J., Nagai, Y., Taniguchi, T., Gomi, H., & Tenenbaum, J. (2021). World model learning and inference. *Neural Networks, 144*, 573–590. https://doi.org/10.1016/j.neunet.2021.09.011

Gao, J., Zheng, P., Jia, Y., Chen, H., Mao, Y., Chen, S., Wang, Yi., Hua, Fu., & Dai, J. (2020). Mental health problems and social media exposure during COVID-19 outbreak. *PLoS ONE, 15*(4), e0231924.

Gillath, O., Abumusab, S., Ai, T., Branicky, M. S., Davison, R. B., Rulo, M., Symons, J., & Thomas, G. (2023). How deep is AI's love? Understanding relational AI. *Behavioral and Brain Sciences, 46*, e33.

Glock, H.-J. (2019). Agency, intelligence and reasons in animals. *Philosophy, 94*(04), 645–671. https://doi.org/10.1017/S0031819119000275

Glock, H.-J. (2009). Can animals act for reasons? *Inquiry, 52*(3), 232–254. https://doi.org/10.1080/00201740902917127

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*(1), 19–29. https://doi.org/10.1007/s10676-008-9167-5

Holohan, M., Buyx, A., & Fiske, A. (2023). Staying curious with conversational AI in psychotherapy. *The American Journal of Bioethics, 23*(5), 14–16. https://doi.org/10.1080/15265161.2023.2191059

Jackson, R. B., & Tom W. (2021). "A theory of social agency for human-robot interaction." *Frontiers in Robotics and AI* 8 (August 13, 2021): 687726. https://doi.org/10.3389/frobt.2021.687726

Jecker, N. S. (2023). Social robots for later life: Carebots, Friendbots and Sexbots. In R. Fan & M. J. Cherry (Eds.), *Sex Robots: Social Impact and the Future of Human Relations* (pp. 20–40). Springer.

Jecker, N. S. (2021). Nothing to be ashamed of: Sex robots for older adults with disabilities. *Journal of Medical Ethics, 47*(1), 26–32. https://doi.org/10.1136/medethics-2020-106645

Jozuka, E., Sato, H., Chan, A., & Mulholland, T. (2018). "Beyond dimensions: The man who marries a hologram." *CNN*, December 29, 2018. https://www.cnn.com/2018/12/28/health/rise-of-digisexuals-intl/index.html

Karaian, L. (2022). "Plastic fantastic: Sex robots and/as sexual fantasy." *Sexualities*, June, 136346072211066. https://doi.org/10.1177/13634607221106667

Khan, R., & Das, A. (2018). *Build better chatbots: A complete guide to getting started with chatbots*. Springer.

Levy, D. N. L. (2007). *Love + sex with robots: The evolution of human-robot relations* (1st ed.). HarperCollins.

Lewis-Martin, J. (2022). What kinds of groups are group agents? *Synthese, 200*(4), 283. https://doi.org/10.1007/s11229-022-03766-z

Lingen, V., Marlies, N., Noor, A. A., Giesbertz, J. P., Tintelen, V., & Jongsma, K. R. (2023). Why we should understand conversational AI as a tool. *The American Journal of Bioethics, 23*(5), 22–24. https://doi.org/10.1080/15265161.2023.2191039

List, C. (2021). Group agency and artificial intelligence. *Philosophy & Technology, 34*, 1213–1242.

Ma, J., Tojib, D., & Tsarenko, Y. (2022). Sex robots: Are we ready for them? An exploration of the psychological mechanisms underlying people's receptiveness of sex robots. *Journal of Business Ethics, 178*(4), 1091–1107.

Marečková, A., Androvičová, R., Bártová, K., Krejčová, L., & Klapilová, K. (2022). Men with paraphilic interests and their desire to interact with a sex robot. *Journal of Future Robot Life, 3*(1), 39–48. https://doi.org/10.3233/FRL-210010

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Mecacci, G., Calvert, S. C., & Sio, F. S. D. (2023). Human–machine coordination in mixed traffic as a problem of meaningful human control. *AI & Society, 38*(3), 1151–1166. https://doi.org/10.1007/s00146-022-01605-w

Natale, S. (2021). *Deceitful media: Artificial intelligence and social life after the Turing test*. Oxford University Press.

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics, 24*(4), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Nyholm, S. (2020). Human-robot collaborations and responsibility-loci. In *Humans and Robots: Ethics, Agency, and Anthropomorphism*. Philosophy, Technology and Society. London New York: Rowman & Littlefield International.

Nyholm, S. (2023). Tools and/or agents? Reflections on Sedlakova and Trachsel's discussion of conversational artificial intelligence. *The American Journal of Bioethics, 23*(5), 17–19. https://doi.org/10.1080/15265161.2023.2191053

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative agents: Interactive simulacra of human behavior* (http://arxiv.org/abs/2304.03442). arXiv. http://arxiv.org/abs/2304.03442

Paul, S. K. (2021). *Philosophy of action: A contemporary introduction*. Routledge Contemporary Introductions to Philosophy. New York London: Routledge, Taylor & Francis Group.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.

Schlosser, M. (2019). Agency" *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2019/entries/agency/

Schwitzgebel, E., & Shevlin, H. (2023, March 5). Opinion: Is it time to start considering personhood rights for AI chatbots? *Los Angeles Times*. https://www.latimes.com/opinion/story/2023-03-05/chatgpt-ai-feelings-consciousness-rights

Shanahan, M. (2023). "Talking about large language models." arXiv, February 16, 2023. http://arxiv.org/abs/2212.03551.

Sparrow, R. (2021). Sex robot fantasies. *Journal of Medical Ethics, 47*(1), 33–34. https://doi.org/10.1136/medethics-2020-106932

Sternlicht, A. (2023). CarynAI will be your girlfriend for $1 a minute. *Fortune.* https://fortune.com/2023/05/09/snapchat-influencer-launches-carynai-virtualgirlfriend-bot-openai-gpt4/ (visited on August 7,2023).

Steward, H. (2009). Animal agency. *Inquiry, 52*(3), 217–31. https://doi.org/10.1080/00201740902917119

Strasser, A. (2022). Distributed responsibility in human–machine interactions. *AI and Ethics, 2*(3), 523–532. https://doi.org/10.1007/s43681-021-00109-5

Swanepoel, D. (2021). Does artificial intelligence have agency?. *The mind-technology problem: Investigating minds, selves and 21st century artefacts,* 83–104.

Symons, J. (2001). *On Dennett*. Wadsworth.

Symons, J. (2010). The individuality of artifacts and organisms. *History and philosophy of the life sciences*, 233–246.

Symons, J., & Alvarado, R. (2022). Epistemic injustice and data science technologies. *Synthese, 200*(2), 87.

Symons, J., & Elmer, S. (2022). Resilient institutions and social norms: Some notes on ongoing theoretical and empirical research. *Merrill Series on The Research Mission of Public Universities*.

The Sparrow Team. (2022). Training an AI to communicate in a way that's more helpful, correct, and harmless. *Building Safer Dialogue Agents*. Retrieved March 10, 2023, from https://www.deepmind.com/blog/building-safer-dialogue-agents

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. https://doi.org/10.48550/ARXIV.2302.08399

van Grunsven, J. (2022). Anticipating sex robots: A critique of the sociotechnical vanguard vision of sex robots as 'good companions'. In *Being and value in technology*, pp. 63–91. Cham: Springer International Publishing.

van Hateren, J. H. (2015). The origin of agency, consciousness, and free will. *Phenomenology and the Cognitive Sciences, 14*(4), 979–1000. https://doi.org/10.1007/s11097-014-9396-5

van Hateren, J. H. (2016). Insects have agency but probably not sentience because they lack social bonding. *Animal Sentience* 1, no. 9. https://doi.org/10.51291/2377-7478.1130

Véliz, C. (2021). Moral zombies: Why algorithms are not moral agents. *AI & SOCIETY, 36*(2), 487–497. https://doi.org/10.1007/s00146-021-01189-x

Weber-Guskar, E. (2022). How to feel about emotionalized artificial intelligence? When robot pets, holograms, and chatbots become affective partners. *Ethics and Information Technology, 23*(4), 601–610.

Wooldridge, M., & Nicholas, J. (June 7, 1995). Intelligent agents: Theory and practice." *The Knowledge Engineering Review,* 10(2),115–152. https://doi.org/10.1017/S0269888900008122

Yang, M. (2020). Painful conversations: Therapeutic chatbots and public capacities. *Communication and the Public, 5*(1–2), 35–44. https://doi.org/10.1177/2057047320950636

Zhu, Q. (2020). Ethics, society, and technology: A Confucian role ethics perspective. *Technology in Society, 63*, 101424.