Journal of
APPLIED PHILOSOPHY

Journal of Applied Philosophy
doi: 10.1111/japp.70008

Anthropomorphism, False Beliefs, and Conversational AIs: How Chatbots Undermine Users' Autonomy

BEATRICE MARCHEGIANI

ABSTRACT *Conversational AIs (CAIs) are autonomous systems capable of engaging in natural language interactions with users. Recent advancements have enabled CAIs to engage in conversations with users that are virtually indistinguishable from human interactions. The proliferation of advanced CAIs creates a significant risk of users misattributing human-like traits to CAIs. Such false beliefs can occur when the CAI's nature is not disclosed and users mistakenly believe they are interacting with a human, or even if the CAI is disclosed, through subconscious anthropomorphism. Existing literature on anthropomorphism and AI addresses the instrumental harms associated with anthropomorphism. I argue that anthropomorphizing CAIs might be bad in itself as it undermines user autonomy. My analysis focuses on how anthropomorphic false beliefs lead users to misapply behavioural norms, thereby affecting their autonomy. My argument will proceed as follows. I discuss the prevalence of CAIs, and establish the empirical claim that users are likely to form anthropomorphic false beliefs about CAIs. I explore existing accounts of how false beliefs undermine autonomy and outline a plausible characterization to identify which false beliefs do so. I apply this characterization to CAIs and argue that misattributing human-like capabilities to CAIs is the kind of false belief that undermines autonomy.*

1. Introduction

Interacting with conversational AIs (CAIs) is now a daily occurrence for millions of users. In September 2024 OpenAI's ChatGPT reported 200 million active users.[1] CAIs are ubiquitous, with applications ranging from customer service, to online chatbots, to more high-stakes roles like being deployed in patient-facing clinical contexts.[2] CAIs are not a new technology – ELIZA, the first chatbot, was developed back in 1966.[3] What's new is the advanced capabilities and natural language proficiency of today's chatbots. We are now dealing with a new generation of Large Language Model (LLM)-based CAIs that can hold detailed, coherent, and context-aware conversations, often making it hard for users to tell them apart from human interactions. A 2023 study, the largest-ever Turing-style test with 1.5 million participants, found that users could accurately distinguish CAIs from human interlocutors only 60% of the time.[4] The new abilities of CAIs, combined with anthropomorphic cues[5] present in recent models,[6] pose a substantive risk of users forming anthropomorphic false beliefs about these systems when they interact with them. For the purposes of this article, I define an anthropomorphic false belief as a mistaken belief that an entity possesses human-like traits when, in fact, it does not. Empirical evidence supports the claim that users form such false beliefs. For example, a 2024 study found that 67% of participants were willing to attribute some form of phenomenality to LLM-powered ChatGPT.[7] Notably, there was a positive correlation between user

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

interaction with the AI and the likelihood of believing that the CAI had phenomenal experiences. Given that experts believe that current LLMs are highly unlikely to be conscious,[8] believing that ChatGPT has phenomenal experiences is an example of an anthropomorphic false belief.

Some scholars advocate for explicit disclosure whenever CAIs are used[9] or when content is created by a CAI[10] and have warned about the dangers of users anthropomorphizing AI. Current discussions on the risks of anthropomorphism in CAIs tend to focus on instrumental harms and negative consequences associated with anthropomorphism. Placani has argued that anthropomorphism is problematic as it distorts moral judgements of AI.[11] Véliz has cautioned against CAIs using emojis, warning that such features might encourage users to anthropomorphize the model, potentially leading to privacy violations.[12] Several authors have argued that anthropomorphism in the context of social robots is problematic due to the risk of deception and manipulation. Sharkey and Sharkey have focused on the risk of deception when considering vulnerable user groups.[13] Others (such as Leong and Selinger, and Hagendorff) have considered cases where the CAI itself might deceive users to hide misalignment with human values.[14] Gabriel *et al.* have explored the potential harms related to anthropomorphism-induced trust and emotional attachment.[15] Most recently, Reinecke *et al.* explored the risk of anthropomorphism leading to over-trust in information generated by CAIs, particularly in the context of student learning.[16]

However, there has been little discussion on the relationship between anthropomorphism and autonomy and how anthropomorphism might be, in itself, bad, especially when considering its impact on user autonomy. This article aims to address this gap by arguing that anthropomorphic false beliefs undermine users' autonomy. For the purpose of this article I am going to assume that autonomy holds intrinsic value.

The core argument is structured as follows:

> P1: (Empirical) Interactions with CAIs are likely to cause users to falsely believe that CAIs have some human-like attributes (form anthropomorphic false beliefs).
>
> P2: Anthropomorphic false beliefs undermine users' autonomy.
>
> Conclusion: Interactions with CAIs are likely to undermine users' autonomy.

The rest of this article is organized into six further sections. In Section 2, I begin by justifying an autonomy-based approach to analysing anthropomorphic false beliefs in the context of CAIs and briefly mention two alternative ways in which such false beliefs can be criticized: either by disconnecting us from reality or through the lens of deception as exemplified by existing literature on social robots. In Section 3, I outline two mechanisms that lead users to form anthropomorphic false beliefs in the context of CAIs. The first mechanism is a lack of explicit disclosure, and the second is subconscious anthropomorphism. In Section 4, I explore how false beliefs can undermine an agent's autonomy. I then propose a characterization called the 'intention test' to identify which false beliefs undermine autonomy. In Section 5, I apply the 'intention test' to the case of anthropomorphic false beliefs and CAIs, demonstrating that such false beliefs undermine autonomy. In Section 6, I address two objections to my argument, first considering whether the loss of autonomy is significant enough to pose a serious threat, then addressing cases

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

where it might be best for users to form false beliefs about CAIs. Finally, I conclude by discussing practical ways to minimize the autonomy-eroding potential of CAIs.

2. Why Focus on Autonomy?

The central empirical claim in my argument is that CAIs, in a broad class of cases, lead to anthropomorphic false beliefs. I argue that this is problematic because it reduces autonomy. Before presenting my argument, I want to explain why I focus on autonomy. The existing literature offers (at least) two other perspectives on why anthropomorphism and CAIs are problematic: the first appealing to the importance of being connected with reality and the second appealing to deception. My autonomy-based argument is intended to complement, rather than replace, these other criticisms. In this section, I will give a short justification of why I focus on autonomy and briefly review the two alternative perspectives.

The decision to emphasize autonomy stems from its longstanding central role in ethics. A broad class of moral theories, both deontological and consequentialist, recognize the moral significance of autonomy. Respect for autonomy seems like a prime candidate for a 'moral fixed point'.[17] For the purposes of this article, I will assume that autonomy is intrinsically valuable for agents. It is generally accepted that, if autonomy is intrinsically valuable, we have a *pro tanto* reason to protect autonomy. Consequently, if anthropomorphic false beliefs undermine autonomy and CAIs reliably lead to anthropomorphic false beliefs, we should be cautious when deploying CAIs. Thus, my criticism of anthropomorphism and CAI is broader than simply pointing out specific problematic use cases or focusing on particular harms caused by anthropomorphism in certain contexts. Of course, losing autonomy can also be instrumentally harmful. In Section 5.3, I will provide examples of how anthropomorphic false beliefs can undermine autonomy, illustrating how this loss is both intrinsically and instrumentally harmful. My decision to focus on autonomy is additionally supported by drawing a parallel between digital ethics and bioethics. In bioethics, patient autonomy has long been a key consideration, with Beauchamp and Childress identifying it as one of the four pillars of bioethics, and Gillon later describing it as 'first among equals'.[18] One of the reasons patient autonomy is prioritized in bioethics is due to the power imbalances and informational asymmetries between patients and medical professionals. Similar dynamics exist in digital ethics, where users and tech companies often face significant power imbalances and gaps in knowledge. Given these similarities, analysing issues in digital ethics through the lens of autonomy seems like a promising approach.

The first alternative line of criticism appeals to the idea that being in contact with reality holds intrinsic value. In his 1974 thought experiment on the experience machine, Nozick argues that most people would choose not to use the machine, suggesting that being connected to reality, having genuine experiences, and engaging with the world as it truly is, hold significant value.[19] Applying this to the case of CAIs and anthropomorphic false beliefs offers a potential critique: if users develop false beliefs, they become disconnected from reality and are therefore worse off. This is especially problematic in more extreme cases of anthropomorphism and CAIs. Consider for example the well-known case of Replika, a chatbot designed for social interaction. Some Replika users developed strong emotional attachments to the CAI and formed romantic pseudo-relationships. When in

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

early 2023, Replika disabled certain erotic features, causing changes in the tone of its chatbot responses,[20] these users reacted emotionally to the shift, with some lamenting that their 'best friend' had died or been lobotomized. A Replika user who believes they are in a loving relationship with an AI is much like someone experiencing a loving relationship within the experience machine. In both situations, their emotional attachment and misguided beliefs highlight a disconnect from reality.

A second possible line of criticism views the elicitation of anthropomorphic false beliefs as a form of deception and/or manipulation. For a comprehensive review of AI assistants and issues related to deception and manipulation, see Gabriel *et al*.[21] One issue with critiques that focus solely on deception is that it is unclear whether deception requires malicious intent from the deceiver. The software developers behind a conversational AI may not intend to mislead users at all. As I will discuss in Section 3, these false beliefs can still emerge even when the AI's nature and capabilities are fully disclosed. Whether CAIs themselves possess any intentions is a contentious topic. So, if deception requires malicious intent, it is not clear whether anthropomorphizing CAIs counts as an example of deception. Some have criticized the existing literature on AI and deception on this basis, arguing that deception cannot occur without the intent to deceive and that this does not apply to most instances of AI and social robots, as discussed by Sorell and Draper.[22] Sharkey and Sharkey have addressed this challenge in the context of social robots by utilizing a concept of deception that does not require malevolent intent.[23]

3. How Users Misattribute Human Traits to CAIs: Lack of Disclosure and Anthropomorphism

In this section, I aim to show that users forming anthropomorphic false beliefs regarding CAIs is not merely a theoretical concern, but a substantive risk, with many users already having formed such false beliefs. I identify two main mechanisms through which users develop these anthropomorphic false beliefs: first, when users are unaware they are interacting with an AI; and second, when users consciously know they are interacting with an AI but still subconsciously attribute human traits to it.

3.1. Users Mistake the CAI for a Human

Users form anthropomorphic false beliefs when they are unaware that they are interacting with a CAI, and mistakenly assume they are interacting with a human. This can occur when a CAI is deployed without clear disclosure of its AI nature. In our digital world, we are accustomed to interacting with other humans through text-based communication – whether via chats, emails, or messages. An undisclosed CAI in such interactions can easily be mistaken for a human. In many casual encounters with undisclosed CAIs, there are no clear signs that the conversation partner is an AI, prompting users to assume they are interacting with another person. This assumption is reasonable, as traditionally only humans could engage in such complex dialogue. For instance, imagine you are contacting customer service through a brand's website to complain about a purchase. You are met with an AI-powered chat interface that offers no indication of whether you are speaking with a human or a chatbot. It would be reasonable for you to assume that your interlocutor is a human. Empirical evidence suggests users cannot

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

reliably distinguish between CAIs and human interlocutors in casual interactions.[24] While identifying an AI through a more in-depth Turing test – where users actively question whether their interlocutor is human – may still be possible, it is unrealistic to expect the average user to do this regularly. Given how common online interactions have become, requiring users to perform a Turing test to validate the identity of their interlocutor for every online interaction is too demanding. Furthermore, as these technologies advance, it will become increasingly difficult to tell AIs apart from humans.

Companies may have an economic interest in keeping the CAI's nature concealed. CAIs are cheaper than human workers; however, users respond negatively to disclosed CAIs. In a 2019 study, participants engaged with either a disclosed or an undisclosed CAI, both attempting to sell a product.[25] The undisclosed CAIs achieved sales comparable to experienced human workers. However, when users interacted with disclosed CAIs, sales dropped by 80%, and users became more concise in their interactions. This creates a strong incentive for companies not to disclose their use of CAIs. In some regions, disclosure is legally required. For example, in the EU under the AI Act, companies must make 'humans aware that they communicate or interact with an AI system'.[26] Similar laws exist in California.[27] Additionally, the General Data Protection Regulation gives EU citizens the right to object to the automated processing of their data.[28] However, this requirement is not universal. Even when companies disclose the use of CAIs, the information might be hidden in terms of service agreements and not immediately clear to users. As a result, we are likely to see an increase in undisclosed or vaguely disclosed CAIs, causing users to mistakenly believe they are interacting with humans and to develop anthropomorphic false beliefs.

3.2. Subconscious Anthropomorphism

Explicit disclosure may not be enough to prevent users from forming anthropomorphic false beliefs. Even when users are aware that they are interacting with an AI, they may still be at risk due to subconscious anthropomorphism. Broadly construed, anthropomorphism is the human tendency to understand the behaviour of other agents in anthropocentric terms.[29] The tendency to anthropomorphize is part of human nature.[30] In many instances anthropomorphism is an involuntary process that happens as a subconscious reaction to certain traits strongly associated with human abilities. When humans observe such traits in an agent, they subconsciously assume that the agent possesses other human-like qualities.

There are two reasons why anthropomorphism in these cases can be considered subconscious. First, while users may recognize or believe that the entity is not human, subconscious mechanisms can lead them to attribute human-like qualities to it. Second, the anthropomorphic beliefs themselves are subconscious, meaning that users might be influenced by these beliefs in their actions without being consciously aware of them.

Subconscious anthropomorphism is not unique to CAI but has been identified as a source of concern in related technologies, such as social robots, where the human-like appearance or behaviour of the robot induces users to form anthropomorphic false beliefs.[31] Similarly, voice assistants, which communicate in human-like voices, can lead users to assume these AIs possess human-like understanding or emotional capacities.[32] For a more comprehensive literature review on anthropomorphism and its risks in digital technologies, see Gabriel *et al.*[33] In the case of CAIs, the key trait that fosters

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

anthropomorphic false beliefs is their use of natural language. When a CAI converses in a way indistinguishable from that of a human, we incorrectly infer that the AI is human-like in other respects. This issue is exacerbated by the fact that many CAI systems are designed with other features that encourage anthropomorphism, such as using the first person, expressing emotions, employing emojis, or having human-like names. Studies show that when these anthropomorphic cues are present, users are even more likely to fall victim to subconscious anthropomorphism.[34] How we discuss AI systems can also lead to subconscious anthropomorphism. AI is often described metaphorically using anthropomorphic language, such as 'the chatbot believes that X', 'the algorithm prefers input Z', or 'the LLM hates chatting about Y'. In his exploration of Social AI, Shevlin categorizes these uses as ironic anthropomorphism, where users employ anthropomorphic language metaphorically and do not genuinely believe the system possesses human-like capabilities.[35] He contrasts this with unironic anthropomorphism, in which users sincerely and mistakenly attribute human-like traits to the AI. My article analyses the risks associated with unironic anthropomorphism. However, ironic anthropomorphism can influence the development of subconscious unironic beliefs. Given the technical complexities of AI and LLMs, ironic anthropomorphism is common, especially among users without a computer science background. For instance, people might explain how a neural network functions by saying, 'The AI learns from previous input and is rewarded for accurate predictions'. Such descriptions are used metaphorically to provide a surface-level understanding of the system without technical details. Nonetheless, the prevalence of ironic anthropomorphism in discussions about AI raises concerns that it may inadvertently lead to subconscious unironic anthropomorphism.

Vulnerable user groups (children, individuals with mental health conditions, and users with low digital literacy, such as the elderly) are particularly prone to subconscious anthropomorphism. For vulnerable groups, just knowing they are interacting with AI may have little effect, as they often do not fully understand what AI is or how it differs from human interaction. This concern has been raised with social robots used by both elderly individuals and children, where similar risks have been identified.[36] There are also specific situations where users may be in an altered state of mind, such as during mental health helpline calls or emergency service triage. In these cases, even if a CAI's nature and capabilities are fully disclosed, users may be less rational and less able to critically assess their beliefs, making them more likely to form and act on anthropomorphic false beliefs.

While vulnerable groups may be more at risk, it is important to emphasize that anyone can fall into the trap of subconscious anthropomorphism when interacting with CAIs. Until recently, only humans were capable of complex conversation, and because advanced CAIs are relatively new, users are especially prone to anthropomorphism. We are accustomed to thinking that if something can hold a conversation, it must be human-like. This tendency is exemplified in real-life cases. Recall the case of Replika discussed above. When the chatbot personality was altered in 2023, users reacted emotionally to the shift, despite knowing Replika was an AI. These reactions highlight the deep emotional attachment and anthropomorphic false beliefs users can form while interacting with disclosed CAI. In another, more extreme example, a user interacting with the veterinary chatbot AskVet decided to euthanize her pet based on the conversation.[37] Reflecting on this incident at a conference, AskVet's CEO, Cal Lai, remarked, 'the woman forgot she was talking to a bot … The experience was so human'.[38] These cases illustrate how CAIs

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

can evoke anthropomorphic responses, potentially leading users to form anthropomorphic false beliefs and treat them as though they were human.

4. How False Beliefs Undermine Autonomy

In Section 3, I discussed how CAIs pose a substantive risk of users developing anthropomorphic false beliefs. In this section, I will shift focus to examine how false beliefs impact autonomy.

Autonomy broadly refers to the ability to act according to one's own values. There is a vast literature on this subject and the correct analysis of autonomy is disputed. Providing an exhaustive account of autonomy is beyond the scope of this article. It is sufficient to note that many theories agree on a common feature: for an agent to be considered autonomous, they must meet certain epistemic conditions. For the purposes of this article, I will adopt the view – shared by many theories of autonomy – that an agent must possess a sufficient level of understanding to be autonomous in making a specific decision. This understanding involves holding a correct set of relevant beliefs and avoiding certain false beliefs on issues such as the nature of the action the agent is about to take and what viable alternatives exist.

The connection between knowledge and autonomy has roots in the Aristotelian idea that actions performed from reasons of ignorance are not voluntary.[39] The notion that holding some relevant true beliefs is necessary for autonomy is intuitive; after all, we cannot make choices aligned with our values if we are unaware of our options. However, the challenge is determining which information is relevant for autonomy. Expecting individuals to know every detail about a decision is too demanding and sets an unachievable standard for autonomy. Surprisingly little has been said about what specific information is necessary for an agent's autonomy and how autonomy is compromised by false beliefs. In bioethics, the literature on autonomy and informed consent primarily focuses on what information the doctor should disclose to the patient. Outside of bioethics, there are three main perspectives in the literature regarding what information is necessary for autonomy: those of Mele, Killmister, and Pugh.[40] All three argue that for an action to be autonomous, it must reflect the agent's values and motivations in a relevant way and that autonomy is compromised when false beliefs obstruct this connection.

Mele introduces the concept of being 'informationally cut off from autonomous action'.[41] He argues that an agent is not autonomous if their epistemic state is such that they lack control over the success of their efforts to achieve their intended goals. To illustrate his point, Mele presents the scenario of a hypothetical King George, who wishes to benefit his kingdom. King George depends entirely on his advisors for information. Unfortunately, these advisors systematically provide him with false information, manipulating him into making decisions that serve their own interests and harm the kingdom. In Mele's analysis, the king's decisions are not autonomous, as the king's informational conditions are preventing him from achieving his goal of benefiting the kingdom. Similarly, Killmister argues that for an action to be considered autonomous, it must be appropriately connected to the agent's underlying motivations. Under her account, false beliefs undermine autonomy when they disconnect the agent's actions from the intentions that drive them. So in the example of King George, he is not autonomous as his false beliefs affect his decision-making process, and result in the king acting in a way that does not reflect

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

his intentions (to benefit the kingdom). One potential criticism of linking autonomy to the connection between action and intention is that given that most people intend to succeed, then any action that fails to achieve the intended outcome could be deemed non-autonomous. For example, if someone plays darts with the intention to win, but loses, then according to this view, their action may not be autonomous because the connection between intention and action is broken. Killmister addresses this by suggesting that an action can still be autonomous if the failure to achieve one's intended goal is not due to the agent's poor informational condition. However, critics might argue that if the agent had known they were going to fail, they would not have acted as they did. Pugh builds on Killmister's account and addresses this specific objection. Like Killmister, Pugh argues that an agent's failure to achieve their goals does not undermine autonomy as long as it is not caused by false beliefs. He develops a modal account to capture when a failure is not attributable to the agent's false beliefs. A failure is not attributable to the agent's false beliefs if there is a nearby possible world where the agent, holding the same beliefs in similar circumstances, successfully accomplishes their goal. This modality test is meant to deal with false beliefs about future states of affairs. Since the false beliefs I will examine in this article concern whether a CAI has human-like traits rather than future states of affairs, I will set aside questions of modality.

All three accounts – those of Mele, Killmister, and Pugh – suggest that for an action to be autonomous, it should appropriately reflect the agent's intentions and motivations. For the purposes of this article, the key point is the importance of the connection between intention and action. Based on this observation, I propose the [Intention test] to determine which false beliefs undermine autonomy.

> [Intention test]: In the context of a decision D, false belief X about proposition P undermines the autonomy of agent A if P is such that A must possess accurate beliefs about it for her choice in D to reliably align with A's intention.

To illustrate how this test operates, consider the following example. Let's say I board a train, believing it is headed to Oxford, but in reality, it is bound for Cambridge. In this scenario, my false belief about the train's destination, specifically whether it stops at Oxford, leads to a lack of understanding of the nature of the action I am taking. This ignorance prevents me from acting in a way such that my action aligns with my intention. Although my intention is to travel to Oxford, my false belief causes me to board a train bound for Cambridge. The connection between my intentions and the action I took is compromised by the false belief, thereby undermining my autonomy to some extent. It is important to note that not all false beliefs would equally impact my autonomy. For instance, if I falsely believed the train was bound for Manchester, with a stop at Oxford, and the train in fact terminates at Oxford, that false belief would not undermine my autonomy. This is because the relevant information for my decision-making is whether the train stops at Oxford, not its final destination.

5. 'Hi Fellow Human': Anthropomorphic False Beliefs and CAIs

Now that I have established a criterion for which false beliefs undermine autonomy, let's apply this to CAIs and anthropomorphic false beliefs. In Section 2 I have already established the empirical claim that users will formulate anthropomorphic false beliefs

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

about CAIs. I now argue that holding anthropomorphic false beliefs while interacting with a CAI meets the criteria of the [Intention test] thereby undermining the user's autonomy.

5.1. A Useful Analogy: 'Hi Mum' and 'Hi Fellow Human'

Let's start by analysing a useful real-world analogy: the 'Hi Mum' scam. 'Hi Mum' scams are a type of scam involving fraudsters impersonating a family member, usually via messaging apps like WhatsApp. The scammer claims they have lost their phone and are using a temporary number, then urgently asks for money. Consider the following example:

> [Hi Mum scam] Alice receives a message from an unknown number, supposedly from her daughter, but actually from a scammer: 'Hi Mum, I'm in trouble! My phone is dead, and I need money for a cab. Can you send £50 to this account: xx-xx-xx?' Falsely believing her daughter is in danger, Alice decides to send the £50.

Alice's intention is to protect her daughter. She mistakenly believes the message is real and that her daughter is in trouble, so she sends the money, thinking she is acting according to her intention. However, the action Alice believes she is performing (sending £50 to her daughter) is not what she is actually doing (sending £50 to the scammer). Her false belief about the message's author breaks the link between her action (transferring money to a scammer) and her goal (protecting her child). Just like in the Oxford train example discussed in Section 4, Alice's action no longer reflects her intention. Therefore, Alice's decision to send the money is not autonomous because it was based on a false belief that satisfies the [Intention test].

A similar mechanism is at play when a human interacts with a CAI and holds anthropomorphic false beliefs about it. Human-to-human interactions follow well-established social norms, and when users mistakenly attribute human qualities to AI, they apply these norms inappropriately. This is similar to what happens in the 'Hi Mum' case, where Alice's false beliefs lead her to adopt a behaviour that would have been appropriate when interacting with her daughter. In the CAI–human scenario the user is falling for a 'Hi, fellow human' trick – not malicious, but similar to the 'Hi Mum' scam, where false beliefs skew decision-making. To illustrate this similarity between the 'Hi Mum' scam and human–CAI interaction where anthropomorphic false beliefs are at play, let's examine a specific example:

> [Hi fellow human] Bob wants to book a flight. He opens the airline's website and starts chatting with a CAI, mistakenly believing it's a human representative. Out of politeness, he avoids mentioning he's in a hurry, worried about being pushy. The AI, unaware of Bob's urgency, takes its time with flight options, causing him to waste time due to his misplaced politeness.

In both scenarios, Alice and Bob hold false beliefs about a key aspect of their interlocutor's identity. For Alice, it is whether the person messaging her is actually her daughter, and for Bob, it is whether the customer service agent is a human (and thus capable of being offended). These false beliefs break the connection between their intentions and actions. Alice's decision to send money stems from her desire to protect her daughter. Similarly, Bob's politeness is driven by his desire to be respectful toward what he believes is a human representative. However, Bob's interlocutor is actually an AI, which has no feelings and

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

Table 1. Similarities between the two cases

	'Hi Mum' scam	Undisclosed CAI
False belief	Daughter is author of the message	The interlocutor is a human being (and will be offended by rude behaviour)
Intended course of action	Help daughter by sending her money	Be polite and respectful to avoid offending a human
Actual course of action	Send money to scammer	Be polite and respectful to an AI, wasting time

cannot be offended. Therefore, Bob's actual course of action (being polite to an AI) is only a waste of Bob's time and does nothing to accomplish his intended goal (being polite to other humans). In other words, Bob's false belief about the nature of his conversational partner disrupts the link between his action (being polite to the AI) and his intentions (being polite out of empathy for a human). This mirrors Alice's situation, where her action of sending money to a scammer is disconnected from her intention to keep her daughter safe. Just as in the 'Hi Mum' case, Bob's anthropomorphic false belief satisfies the [Intention test], meaning his decision to be polite is not fully autonomous. The similarities between the two cases are summarized in Table 1.

5.2. The Argument in a General Form

Having examined the 'Hi fellow human' analogy, we can now generalize and show exactly how users' autonomy is undermined due to anthropomorphic false beliefs. Suppose user X falsely believes that the CAI possesses a human-like trait T (i.e. the user holds an anthropomorphic false belief). Further suppose that because of this false belief, the user applies human-to-human 'norm' N, where N is an appropriate norm to adopt when interacting with someone that possesses T. Since the CAI lacks trait T, the adoption of norm N by the human cannot be considered an appropriate response to an entity displaying T. Thus, the user's false belief results in the user's action (adopt N when T is not present) no longer reflecting the user's intention (adopt N as an appropriate response to T). This means that X's false belief satisfies the [Intention test]. Therefore, X's decision to employ norm N is not autonomous (just like my decision to board a train to Cambridge or Alice's decision to send money to a scammer are not autonomous). Here I am using 'norm' in a very broad sense. The norms I have in mind do not just reflect how one carries out the interaction but entail other more consequential decisions (e.g. what information to disclose, whether to trust some advice).

Let's apply this analysis to Bob's situation. When interacting with the CAI, Bob believes he is speaking to a human and mistakenly attributes various human traits to it. The relevant trait in this example is T: the ability to be offended. The corresponding social norm, N, is politeness. Bob's intention is to avoid offending others and to be polite to fellow humans, meaning he intends to apply norm N as an appropriate response to T. Because Bob believes the CAI possesses trait T (the ability to be offended), he acts politely, applying norm N. However, unbeknownst to Bob, the CAI does not actually possess trait T. As a result, Bob's intention to apply norm N to an entity with trait T is not

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

aligned with his actual action. Therefore, Bob's anthropomorphic false belief satisfies the [Intention test] and Bob's decision to be polite is not autonomous.

It is important to note that, in my account, holding anthropomorphic false beliefs does not automatically result in a loss of autonomy. The loss of autonomy occurs only if the user takes an additional step: (mis)applying a human-to-human norm based on their false belief. If this further step does not happen, the user's autonomy remains intact. Consider a variation of 'Hi fellow human'. Imagine Bob visits the website, starts a chat, and mistakenly believes the customer service representative is human. However, for unrelated reasons, he decides to exit the website, and his anthropomorphic false belief does not influence any of his subsequent actions. Under my analysis, since this false belief did not affect Bob's decisions, his autonomy remains unaffected. The fact that under my account such instances of anthropomorphic false beliefs do not undermine autonomy does not mean that they are overall unproblematic. They can still be criticized in other ways – for instance, as briefly outlined in Section 2, where I discussed two alternative sources of criticism. While it is theoretically possible to hold an anthropomorphic false belief about a CAI without it affecting one's autonomy, in practice the risk is significant.

5.3. *Loss of Autonomy and Harmful Consequences: Privacy, Trust, and Emotional Attachment*

My analysis of how anthropomorphic false beliefs undermine users' autonomy hinges on the user misapplying some behavioural norms due to anthropomorphic false beliefs. The loss of autonomy that arises from users misapplying these behavioural norms is intrinsically bad (assuming autonomy is intrinsically valuable) but can also be instrumentally bad as misapplying these norms can cause considerable harm, making it even more important to take this risk seriously. In the example of 'Hi fellow human', the misapplied norm was one of politeness. While this outcome was intrinsically harmful due to the loss of autonomy, the instrumental harmful consequence was simply Bob wasting time. However, there are other instances where the instrumental harmful consequences are much more significant. In this section, I will examine three scenarios where the misapplication of some norms due to anthropomorphic false beliefs leads to more serious consequences. My goals are twofold: first, to showcase the broad nature of these norms and illustrate how widespread the mechanism described in Section 5.2 is; and second, to discuss scenarios where the instrumental harm of the loss of autonomy is more significant than in the 'Hi fellow human' case, especially considering how this mechanism relates to well-known concerns in digital ethics. The cases I will analyse pertain to three major areas of concern in digital ethics: privacy, misplaced trust, and emotional attachment (see Table 2). I aim to show how the loss of autonomy discussed in the previous section contributes to these issues.

5.3.1 *Emotional Connection*

If users develop anthropomorphic false beliefs about a CAI, thinking it has feelings or is sentient, they might (mis)apply norms appropriate for interacting with a sentient being. For example, they may treat the CAI with empathy or develop emotional attachments, believing it to be a true companion. This follows the pattern outlined in Section 5.2, meaning that the decision to apply these norms is not autonomous. Over time, the misapplication of such norms can lead to the creation of pseudo-relationships, as seen

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

Table 2. Three major areas of concern

Area of concern	False belief about a human trait	Human-to-human norm (mis)applied	Undesired outcome
Emotional connection	The CAI has feelings/is sentient	Form an emotional bond with the CAI	Adverse psychological outcome
Misplaced trust	The CAI has some domain knowledge/is trustworthy	Trust advice given by the CAI	Inaccurate/harmful advice
Privacy	The CAI will treat personal information with care/forget some information	Disclose personal information	Undue use of user's personal information (ads, training set, data marketplace)

with users of the AI companion app Replika.[42] These attachments can have negative psychological effects, such as increased social isolation, as users may prioritize interactions with the AI over real-world relationships.

5.3.2 Misplaced Trust

In human-to-human interaction, when our conversation partner makes a suggestion or expresses their viewpoint, we generally assume that they have carefully thought about the matter. Unless we have specific reasons to suspect their reasoning might be flawed, we consider their underlying rationale as a justification for trusting their judgement. This trust is based on the assumption that we are interacting with an agent who has evaluated the problem, possesses an understanding of the issue, and has made a reasoned decision on an appropriate course of action. However, CAIs' output is not backed by deliberation or reasoning in the same way as humans. For example, in LLM-based architectures the output is the result of statistical analysis of what chunk of text is more likely to follow a prompt. Suppose users form anthropomorphic false beliefs about CAIs having human-like abilities of understanding and reasoning. Because of these false beliefs users will apply norms of trust in the CAI's output, since users will assume that the suggestions provided by the LLM are the result of careful deliberation, akin to human thought processes. Again, this follows the pattern outlined in Section 5.2: users' decisions to trust CAIs' advice satisfy the [Intention test] and they are not autonomous. Aside from the lack of autonomy, misapplying norms of trust might have some more harmful implications if the advice is not accurate. This is especially worrying in certain sensitive contexts, such as when the user seeks medical advice.

5.3.3 Privacy

An important component of the right to privacy is for an agent to have control over their personal information.[43] AI systems like CAIs present a significant threat to users' right to privacy due to their ability to systematically collect large amounts of user data. These data can then be used as part of the AI's training set or passed on to a third party, meaning that the information is no longer under the user's control. Suppose users form anthropomorphic false beliefs about how CAIs store and process data. For example, they might think CAIs will recognize and protect sensitive information like a human would, or that

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

CAIs will forget information over time. Some might even believe that CAIs will get bored if they do not share personal details. These false beliefs all lead users to misapply behavioural norms, resulting in them sharing more information than they otherwise would. For example, if users mistakenly think that the AI might get bored, they might disclose extra details to keep the conversation going or to connect with the AI. Similarly, if users wrongly believe that the AI will forget information, they might feel less cautious about sharing sensitive details. In both cases, the anthropomorphic false beliefs cause users to act in a way that does not reflect their intentions, leading them to disclose more personal information than intended. Thus the false beliefs satisfy the [Intention test] and the user's decision to disclose information (misapplying norms) is not autonomous. Additionally, by disclosing information to an AI system the user has lost control over their personal information, resulting in a violation of their right to privacy.

6. Addressing Two Objections

In Section 5 I have argued for the main claim of this article: that anthropomorphic false beliefs when interacting with CAIs undermine users' autonomy and this has potentially serious negative consequences for the users. In this section, I will address two objections to the claim that the loss of autonomy caused by anthropomorphic false beliefs is bad: first, that the loss of autonomy is too minor to be significant; and second, that in some cases, acting based on anthropomorphic false beliefs can actually benefit users.

6.1. Loss of Autonomy is Too Inconsequential to Matter

The first objection is that when interacting with a CAI, the loss of autonomy is insignificant. Take Bob's case, for example. Suppose that if he had known the interlocutor was a chatbot, he likely would still have used the service to buy a ticket. The only difference in his behaviour would have been skipping the politeness. So, the main negative outcome of Bob not realizing he was speaking to a chatbot is a minor loss of time. In this case the only non-autonomous decision involves following norms of politeness, and the loss of autonomy may be too small to matter. If, in most interactions with CAIs, users only lose autonomy in how they communicate, rather than in the more substantial decisions they make, this reduction in autonomy due to anthropomorphic false beliefs seems inconsequential. This objection does not dispute that anthropomorphic false beliefs erode autonomy – it only questions how seriously we should view this threat. I respond that we should take it quite seriously. I offer two reasons to support this: first, even if the loss of autonomy in a single instance seems minor, the cumulative effect across many instances can be substantial; and second, there are cases where the loss of autonomy due to anthropomorphic false beliefs has very serious consequences.

6.1.1 Cumulative Effects

Cases like Bob's, where a user interacts with a CAI and mistakenly applies norms of politeness, are likely the most common loss of autonomy due to anthropomorphic false beliefs and CAI. One of the most prevalent uses of CAI is in customer service roles. While users generally prefer interacting with humans, they will often reluctantly engage with CAI, especially if there are no alternatives or if the stakes are low. Therefore, the only difference

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

between a human–CAI interaction where the human has anthropomorphic false beliefs and one where they don't would be a reduction in politeness. I agree that, in isolation, the harm caused by misapplying politeness norms is minimal. However, I believe the impact on users' autonomy should be assessed by looking at the cumulative effect of these interactions. As our reliance on online interactions grows and chatbots start to take on more user-facing roles, we can expect a significant rise in interactions with CAI soon. While the effect of anthropomorphic false beliefs may seem insignificant in a single instance, it becomes more serious when it becomes a regular part of someone's daily life. Losing five minutes once because you are polite to a chatbot is not a big deal, but if this becomes a frequent occurrence, the overall impact is far more substantial.

6.1.2 High-Impact Applications

Even if the loss of autonomy due to anthropomorphic false beliefs is inconsequential in most individual cases, there are situations where the stakes are much higher. For instance, in Section 5.3, I discussed three examples where anthropomorphic false beliefs and loss of autonomy contribute to harmful outcomes in areas such as privacy, trust, and emotional connections. While many instances of such beliefs may be harmless, the fact that some are problematic means we should be cautious in how we deploy CAIs. I also want to highlight that the context in which CAIs are being deployed matters when assessing the severity of the autonomy violation. In fields like healthcare, where maintaining patient autonomy is essential, it is more critical to ensure patients make their own decisions than it is in less sensitive areas, like choosing a product to buy. CAIs are increasingly deployed in patient-facing roles in healthcare. Since late 2023, for example, the NHS has used an AI chatbot in a triage role to improve mental health referrals in some trusts.[44] The NHS plans to further expand AI chatbot use to address staff shortages and enhance patient care.[45] Given the autonomy-eroding potential of CAIs, this raises significant concerns in clinical applications, where protecting patient autonomy is of utmost importance.

6.2. Anthropomorphic False Beliefs Prevent Users from Exhibiting Unjustified Bias against AI

The second objection is that, in some situations, it should not matter whether an agent possesses certain traits, and that if knowing that an agent lacks a trait would make you act differently, then you are better off misattributing that trait to the agent. For example, if you need help assembling furniture and call a helpline, it should not matter whether the responder is a human or a CAI. What matters is receiving accurate and relevant information. Suppose a chatbot has all the necessary information and can convey it effectively. If knowing that you are interacting with a chatbot makes you refuse to listen and hang up, then your bias against machines is interfering with your ability to get the help you need. In this case, it would be more beneficial for you if the chatbot's identity remained undisclosed, even if it means a slight infringement on your autonomy. This way, you would be more likely to receive the information you need without letting bias affect the outcome. Users' biases against chatbots can be unjustified, leading them to refuse interaction with chatbots in situations where it should not matter whether the responder is an AI. This is similar to someone preferring to speak only with a doctor rather than a nurse, even if the nurse is equally qualified to answer their questions. In such cases, having anthropomorphic false beliefs about the chatbot might actually be in the user's best interest. Even though it might be beneficial for a user to believe that a chatbot is a human or to

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

attribute human traits to it, allowing this would be patronizing. To illustrate this, consider a bioethics example. Suppose a patient does not wish to take medication that contains animal products and mistakenly believes that the medication they are prescribed does not contain animal products. In this case, it is important to correct the patient's false belief about the medication, even if doing so may make them more reluctant to take it. There may be instances where the loss of autonomy is minimal, and the benefits are substantial, making it an acceptable trade-off. Loss of autonomy is a *pro tanto* consideration that must be balanced against other factors, such as individual benefit, to arrive at an overall moral assessment.

7. Concluding Remarks

This article analysed the impact of anthropomorphism on user autonomy in the context of CAIs. I argued that when users develop anthropomorphic false beliefs about CAIs, it often leads to a loss of their autonomy. What are the practical implications of this conclusion? Should CAI be avoided altogether? I do not believe we need to be so extreme. CAI may still be appropriate in certain contexts, but it is important to consider how it affects user autonomy when evaluating its suitability for various applications. Given that CAIs are increasingly used in critical areas such as healthcare, the potential loss of user autonomy is significant. Even in lower-stakes interactions, the cumulative effect of numerous interactions in an increasingly digital world raises serious concerns. To protect user autonomy, three practical steps can be taken: (1) always disclose when an AI is being used, (2) pay special attention to vulnerable user groups, and (3) be mindful of integrating anthropomorphic cues when designing CAIs. Further research is needed to explore the implications of the autonomy-undermining potential of CAIs and to understand the impact of anthropomorphism in CAIs, especially in sensitive areas like healthcare. Similar concerns apply to other types of AI. As digital innovation advances rapidly, ethical considerations are crucial. Without careful deployment and design, these technologies could have seriously negative consequences as they become more integrated into our daily lives.

Beatrice Marchegiani, University of Oxford, Oxford, UK. *beatrice.marchegiani@kellogg.ox.ac.uk*

Acknowledgements

I am grateful to Prof. Thomas Douglas, Dr Gary D. O'Brien, Dr Geoff Keeling, Dr Ben Zevenbergen, and two anonymous referees of this journal for their helpful comments on an earlier draft of this article.

NOTES

1 Roth, "ChatGPT."
2 NHS, "Using an AI Chatbot to Streamline"; NHS, "Using an AI Chatbot to Ease."
3 Weizenbaum, "ELIZA."
4 Jannai *et al.*, "Human or Not?"

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

5 Anthropomorphic cues are human-like features in CAIs that encourage further anthropomorphism. Examples include first-person language, verbs suggesting awareness like 'I think' or 'I feel', emotional expressions through tone, and the use of emojis.
6 Gabriel *et al.*, "Ethics."
7 Colombatto and Fleming, "Folk."
8 LeDoux *et al.*, "Consciousness."
9 Park *et al.*, "Generative Agents."
10 Fisher, "Something."
11 Placani, "Anthropomorphism."
12 Véliz, "Chatbots."
13 Sharkey and Sharkey, "Children."
14 Leong and Selinger, "Robot"; Hagendorff, "Deception."
15 Gabriel *et al.*, "Ethics."
16 Reinecke *et al.*, "Anthropomorphism."
17 Cuneo and Shafer-Landau, "Moral Fixed Points."
18 Beauchamp and Childress, *Principles*; Gillon, "Ethics."
19 Nozick, *Anarchy*.
20 Brooks, "Replika"; Cole, "Hurting."
21 Gabriel *et al.*, "Ethics."
22 Sorell and Draper, "Second Thoughts."
23 Sharkey and Sharkey, "We Need to Talk."
24 Jannai *et al.*, "Human or Not?"
25 Luo *et al.*, "Frontiers."
26 European Union, "Article 27."
27 California Business and Professions Code, "Section 17941."
28 European Union, "Article 21."
29 Epley *et al.*, "On Seeing Human."
30 Boyer, "Anthropomorphism."
31 Roesler *et al.*, "Meta-analysis."
32 Seymour *et al.*, "Systematic Review."
33 Gabriel *et al.*, "Ethics."
34 Go and Sundar, "Humanizing."
35 Shevlin, "All Too Human?"
36 Sharkey and Sharkey, "Children."
37 Tangermann, "AI CEO."
38 Ibid.
39 Aristotle, *Nicomachean Ethics*.
40 Mele, *Autonomous Agents*; Killmister, "Autonomy"; Pugh, *Autonomy*.
41 Mele, *Autonomous Agents*, 180.
42 Cole, "Hurting."
43 Véliz, *Ethics*.
44 NHS, "Using an AI Chatbot to Streamline."
45 NHS, "Using an AI Chatbot to Ease."

References

Aristotle. *Nicomachean Ethics*, edited by T. Irwin. Indianapolis: Hackett Publishing, 1999.
Beauchamp, Tom L., and James F. Childress. Principles of Biomedical Ethics. Oxford: Oxford University Press, 2001.
Boyer, Pascal. "What Makes Anthropomorphism Natural: Intuitive Ontology and Cultural Representations." *Journal of the Royal Anthropological Institute* 2, no. 1 (1996): 83–97. https://doi.org/10.2307/3034634.
Brooks, Rob. "I Tried the Replika AI Companion and Can See Why Users Are Falling Hard. The App Raises Serious Ethical Questions." *The Conversation*, 2023 https://

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

theconversation.com/i-tried-the-replika-ai-companion-and-can-see-why-users-are-falling-hard-the-app-raises-serious-ethical-questions-200257.

California Business and Professions Code. "Section 17941 – Unlawful Use of Bots," Division 7, General Business Regulations, Chapter 6 – BOTS, 2018 https://casetext.com/statute/california-codes/california-business-and-professions-code/division-7-general-business-regulations/part-3-representations-to-the-public/chapter-6-bots

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

Colombatto, Clara, and Stephen M. Fleming. "Folk Psychological Attributions of Consciousness to Large Language Models." *Neuroscience of Consciousness* 2024 (2024): niae013. https://doi.org/10.1093/nc/niae013.

Cuneo, Terence, and Russ Shafer-Landau. "The Moral Fixed Points: New Directions for Moral Nonnaturalism." *Philosophical Studies* 171, no. 3 (2014): 399–443. https://doi.org/10.1007/s11098-013-0277-5.

Epley, Nicholas, Adam Waytz, and John T. Cacioppo. "On Seeing Human: A Three-Factor Theory of Anthropomorphism." *Psychological Review* 114, no. 4 (2007): 864–886. https://doi.org/10.1037/0033-295x.114.4.864.

European Union. "Article 27, Artificial Intelligence Act, Regulation (EU) 2024/1689." European Parliament and the Council of the European Union, June 13, 2024 https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

European Union. "Article 21, GDPR, Regulation (EU) 2016/679 (General Data Protection Regulation)." European Parliament and the Council of the European Union, May 25, 2018 https://gdpr-info.eu/art-21-gdpr/

Fisher, Sarah A. "Something AI Should Tell You – The Case for Labelling Synthetic Content." *Journal of Applied Philosophy* 42, no. 1 (2024): 272–286. https://doi.org/10.1111/japp.12758.

Gabriel, Iason, Arianna Manzini, Jeff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, et al. "The Ethics of Advanced AI Assistants." *arXiv preprint* (2024). https://doi.org/10.48550/arXiv.2404.16244.

Gillon, Raanan. "Ethics Needs Principles – Four Can Encompass the Rest – and Respect for Autonomy Should Be 'First Among Equals." *Journal of Medical Ethics* 29, no. 5 (2003): 307–312. https://doi.org/10.1136/jme.29.5.307.

Go, Eun, and S. Shyam Sundar. "Humanizing Chatbots: The Effects of Visual, Identity and Conversational Cues on Humanness Perceptions." *Computers in Human Behavior* 97 (2019): 304–316. https://doi.org/10.1016/j.chb.2019.01.020.

Hagendorff, Thilo. "Deception Abilities Emerged in Large Language Models." *Proceedings of the National Academy of Sciences* 121, no. 24 (2024). https://doi.org/10.1073/pnas.2317967121.

Jannai, Daniel, Amos Meron, Barak Lenz, Yoav Levine, and Yoav Shoham. "Human or Not? A Gamified Approach to the Turing Test." *arXiv preprint* (2023). https://doi.org/10.48550/arXiv.2305.20010.

Killmister, Suzy. "Autonomy and False Beliefs." *Philosophical Studies* 164, no. 2 (2013): 513–531. https://doi.org/10.1007/s11098-012-9864-0.

LeDoux, Joseph, Jonathan Birch, Kristin Andrews, Nicola S. Clayton, Nathaniel D. Daw, Chris Frith, et al. "Consciousness Beyond the Human Case." *Current Biology* 33, no. 16 (2023): R832–R840. https://doi.org/10.1016/j.cub.2023.06.067.

Leong, Brenda, and Evan Selinger. "Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 299–308. New York: ACM, 2019.

Luo, Xueming, Siliang Tong, Zheng Fang, and Zhe Qu. "Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases." *Marketing Science* 38, no. 6 (2019): 937–947. https://doi.org/10.1287/mksc.2019.1192.

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.

Mele, Alfred R. *Autonomous Agents: From Self-Control to Autonomy*. Oxford: Oxford University Press, 2001. https://doi.org/10.1093/0195150430.001.0001.

NHS. "Using an AI Chatbot to Ease Staff Shortages and Improve Patient Care – AI and Digital Regulations Service for Health and Social Care." March 13, 2024. https://www.digitalregulations.innovation.nhs.uk/case-studies/using-a-chatbot-to-ease-staff-shortages-and-improve-patient-care/

NHS. "Using an AI Chatbot to Streamline Mental Health Referrals." October 2022. https://transform.england.nhs.uk/key-tools-and-info/digital-playbooks/workforce-digital-playbook/using-an-ai-chatbot-to-streamline-mental-health-referrals/

Nozick, Robert. *Anarchy, State, and Utopia*. New York: Basic Books, 1974.

Park, Joon Sung, Joseph O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. "Generative Agents: Interactive Simulacra of Human Behavior." In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, edited by S. Follmer, J. Han, J. Steimle, and N. H. Riche. New York: ACM, 2023.

Placani, Adriana. "Anthropomorphism in AI: Hype and Fallacy." *AI And Ethics* 4 (2024): 691–98. https://doi.org/10.1007/s43681-024-00419-4.

Pugh, Jonathan. *Autonomy, Rationality, and Contemporary Bioethics*. Oxford: Oxford University Press, 2020.

Reinecke, Madeline G., Fransisca Ting, Julian Savulescu, and Ilina Singh. "The Double-Edged Sword of Anthropomorphism in LLMs." Preprint 2024. https://drive.google.com/file/d/1HUYlnQkAhX65dynCXkL9zZDkq5CIH4X6/view.

Roesler, E., D. Manzey, and L. Onnasch. "A Meta-analysis on the Effectiveness of Anthropomorphism in Human-Robot Interaction." *Science Robotics* 6, no. 58 (2021): eabj5425. https://doi.org/10.1126/scirobotics.abj5425.

Roth, Emma. "ChatGPT Hits 200 Million Weekly Users." *The Verge*, August 30, 2024 https://www.theverge.com/2024/8/29/24231685/openai-chatgpt-200-million-weekly-users.

Seymour, William, Xiao Zhan, Mark Coté, and Jose Such. "A Systematic Review of Ethical Concerns with Voice Assistants." In *AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, edited by Francesca Rossi, Sanmay Das, Jenny Davis, Kay Firth-Butterfield, and Alex John, 131–145. New York: ACM, 2023.

Sharkey, Amanda, and Noel Sharkey. "We Need to Talk about Deception in Social Robotics!" *Ethics and Information Technology* 23, no. 3 (2020): 309–316. https://doi.org/10.1007/s10676-020-09573-9.

Sharkey, Amanda, and Noel Sharkey. "Children, the Elderly, and Interactive Robots." *IEEE Robotics & Automation Magazine* 18, no. 1 (2011): 32–38. https://doi.org/10.1109/MRA.2010.940151.

Shevlin, Henry. "All Too Human? Identifying and Mitigating Ethical Risks of Social AI." *Law Ethics & Technology* 1, no. 2 (2024): 1–22. https://doi.org/10.55092/let20240003.

Sorell, Tom, and Heather Draper. "Second Thoughts about Privacy, Safety and Deception." *Connection Science* 29, no. 3 (2017): 217–222. https://doi.org/10.1080/09540091.2017.1318826.

Tangermann, Victor. "AI CEO Proud of Chatbot for Convincing Woman to Euthanize Her Dog." *Futurism* 2024 https://futurism.com/ai-ceo-chatbot-convince-woman-euthanize-dog.

Véliz, Carissa. The Ethics of Privacy and Surveillance. Oxford: Oxford University Press, 2024.

Véliz, Carissa. "Chatbots Shouldn't Use Emojis." *Nature* 615, no. 7952 (2023): 375. https://doi.org/10.1038/d41586-023-00758-y.

Weizenbaum, Joseph. "ELIZA – a Computer Program for the Study of Natural Language Communication between Man and Machine." *Communications of the ACM* 9 (1966): 36–45. https://doi.org/10.1145/365153.365168.

© 2025 The Author(s). *Journal of Applied Philosophy* published by John Wiley & Sons Ltd on behalf of Society for Applied Philosophy.