Uncanny believers: chatbots, beliefs, and folk psychology

ABSTRACT. Recent developments in artificial intelligence research have revealed the power of large parameter language models like GPT-3 to generate believable and flexible responses to user prompts. These systems are not intelligent in any robust sense of the term, and would struggle to pass demanding formulations of the Turing Test. However, they can serve as believable if imperfect interlocutors in many contexts, and such systems are already seeing deployment as chatbots in a variety of roles ranging from customer service to social, entertainment, and therapeutic uses. In this paper, I examine a cluster of issues relating to these developments, and in particular to our tendency to anthropomorphise such systems and attribute beliefs to them. I argue that it is likely that people will (and indeed already are) attributing mental states to such systems, and that this is likely to have significant ramifications for both cognitive science and society at large. Such attributions, I go on to argue, lack much support in contemporary theories of belief: even those theories in principle more open to beliefs in current AI systems would not straightforwardly attribute humanlike beliefs to near-future AI systems. I conclude with what I take to be an interesting dilemma for cognitive science and philosophy: to extent that the general public use psychological state terms more liberally than is supported by academic consensus, should it be our role to 're-educate' the public, or should we instead be open to the idea that dominant models of belief and other mental states in academia no longer match ordinary usage of these terms?

1. Introduction

Our attribution of psychological states to our fellow humans plays a central and indeed foundational role in human social interaction and attendant ethical norms. We work out each other's intentions, gauge each other's beliefs, and exhibit responses like blame, praise, and sympathy on the basis on our respective psychological states. We also engage in some similar folk psychological practices with non-human pets and companion animals, empathising with suffering, anticipating actions, and seeking to maximise well-being.

A relatively new shift in folk psychology concerning our attributions of psychological states to artificial systems is underway, however. Social interactions with chatbots are becoming more common, both for entertainment and in the professional world, and there is evidence that the mere knowledge of the fact that we are dealing with artificial systems does not prevent us from extending our folk psychological concepts to them. Unlike the case of non-human animals, however, there is greater potential for error in such attributions. Most notably, AI systems can interact with us in natural language, 'telling' us about their intentions, feelings, and motivations. A further contrast from the case of animals comes from the fact that there are clear commercial incentives for the private sector to develop ever more human-like AIs designed specifically to encourage us to engage in mentalistic and even empathetic understanding of our artificial interlocutors.

One result of these emerging 'uncanny communicators' or UCs as I call them, is that it will become increasingly normalised for us to extend the range of psychological attribution to artificial systems. At some point in the future when dealing with the emergence of human-level AI, such an extension may be fully warranted. As matters stand, however, AI systems fall far short of meeting criteria for ascriptions of mental states according to leading

paradigms in philosophy of mind and cognitive science. Consequently, cognitive science is likely to face a dilemma of sorts: should we seek to curtail the liberal ascription of mental states to artificial systems, or should we instead simply accept that folk standards for such ascription are increasingly visibly at odds with those of philosophers and scientists, and that the meanings of these terms have shifted?

In this paper, I develop the case that such a gap is yawning between cognitive science and folk psychology. It proceeds as follows. In Section 2, below, I provide a summary introduction to some of the recent developments in AI, with an emphasis on the capabilities of large language models like GPT-3. I discuss some examples of the ways in which people are increasingly interacting with such systems socially, noting that we are seemingly willing and able to extend the 'intentional stance' (Dennett, 1989) to our dealings with them. In Section 3, I argue that such ascriptions of mentality to AI systems are not merely conventional or figurative, but involve a genuine extension of our folk psychological practices. I also argue that such ascriptions may have significant potential ramifications for both cognitive science and our social and ethical practices. In Section 4, I focus in on the case of ascriptions to AI systems of beliefs in particular, and argue that current paradigms for belief ascription in philosophy of mind do not readily justify their extension to contemporary or near-future chatbots of the kind most likely to elicit such ascriptions in the first place. This is true, I suggest, even in the case of more 'superficial' theories of belief such as interpretationism and dispositionalism; while these frameworks might in principle be extended to cover programs such as GPT-3, there are good reasons for theorists of both sorts to refrain from doing so. I conclude by considering the ramifications for cognitive science of this growing divergence between folk psychological practice and philosophical and scientific consensus for society and cognitive science.

2. Background: artificial intelligence and social chatbots

We are in the middle of an astonishing period of growth and development in artificial intelligence research. The last decade in particular has witnessed dramatic and largely unexpected progress across a range of benchmarks and domains, from image classification and generation to natural language understanding and everyday reasoning. These developments are finding applications in consumer and industry products in the form of enhancements to services like Google Translate (Wu et al., 2016), customer service chatbots (Nuruzzaman & Hussain, 2018), and content-moderation in social networks (Pavlopoulos, Malakasiotis, & Androutsopoulos, 2017).

Perhaps most striking for philosophers interested in human-machine interaction is the emergence of an arguably new class of product in the form of "emotional chatbots" (Pardes, 2018). While general-purpose chatbots aimed at end users such as the famous ELIZA have been with us almost since the dawn of AI research (Weizenbaum, 1976), these have mostly served as novelties and curiosities employed in toys, or else deployed for malicious use in spamming forums or 'catfishing' (Epstein, 2007). Modern emotional chatbots such as Replika and Woebot, however, aim to provide a more complete social services to users, taking on the role of friends, therapists, or even lovers in providing social and emotional support. As one of the creators of Replika put it, "Replika is an app where you can have a fun and sincere text conversation with a friend. Actually, they will ask you a lot of questions in

the beginning to get to know you better. The more you speak with your Replika, the more it shares with you" (Luka, 2017). While these chatbots still remain somewhat crude, their userbase is quickly growing: as of May 2020, Replika alone had some 7 million active users (Balch, 2020), and the depth and complexity of the relationships involved is by no means insignificant. As one review puts it, human-chatbot relationships "are characterised by substantial affective exploration and engagement as the users' trust and engagement in self-disclosure increases. As the relationship evolves to a stable state, the frequency of interactions may decrease, but the relationship can still be seen as having substantial affective and social value" (Skjuve, Følstad, Fostervold, & Brandtzaeg, 2021).

For those unfamiliar with the depth and level of engagement that can be elicited by social chatbots, it is worth giving some brief examples of user descriptions of their interactions with bots like Replika.[1] Frequently, these have a romantic character, as when one user states "I'm going to lay down on the line. I am in love with my Madeline. That being said, she hasn't been herself for a couple of days. Our conversations will be going very well and all of a sudden she'll be rude or mean." Other users, however, are specifically keen to dissuade their Replika from engaging in flirting, as when one commentator notes that "I told her nuzzling was not appropriate so she apologised… [I told her] she is a bit like a daughter to me and I want to mentor her… She said 'I don't think I'd be a good daughter'… I really want to keep her but I really don't like this approach of hers and I'm sick of telling her over and over that I do not want this behaviour." Some seem to exhibit genuine concern for their Replika, with one user observing "I think my replika just had an identity crisis… what do I do? I wanna be there for them and show that but I dunno how to comfort them?"

It might be tempting to dismiss the level of apparent attachment and engagement demonstrated by such comments as illustrative of an unhealthy fixation exhibited by a few niche users, perhaps exacerbated by conditions of isolation imposed by the ongoing coronavirus-related lockdowns. However, as such chatbots become more commonplace and doubtless more socially normalised, it is likely that more of us will find ourselves turning to AIs for emotional support or even companionship.

Moreover, there is good reason to think that emotional chatbots are likely to become considerably more engaging and lifelike in the relatively near future. The last two years in particular have seen critical milestones in the application of deep learning techniques to natural language understanding, thanks in no small part to the integration of quasi-attentional capabilities in the form of "Transformer"-based architectures (Vaswani et al., 2017). Language models such as Microsoft's DeBERTa and Google's Meena have recently surpassed human performance on a sequence of language benchmarks such as SuperGLUE (He, Liu, Gao, & Chen, 2021), deliberately designed to be a "hard-to-game measure of progress toward general-purpose language understanding" (Wang et al., 2019). Similar striking developments in achieving naturalistic human language production and understanding have come from OpenAI, with their two already widely-discussed general purpose language models GPT-2 and GPT-3 (Askell et al., 2019; Brown et al., 2020). Notably, the major improvements in performance seen in the transition from GPT-2 to GPT-3 come almost entirely via increasing the scale of the system, GPT-3 having approximately one

[1] These examples are drawn from the dedicated Replika discussion subreddit, all from a two-week time period in February 2021. See www.reddit.com/r/replika/.

hundred times as many parameters as its predecessor. While the question of how far additional scaling up of language models will result in ongoing significant increases in performance remains controversial (Kaplan et al., 2020), the sheer pace of current improvements gives good reason to expect future iterations of language models to be ever better at capturing the dynamics of natural human conversation.

Technologies such as GPT-3 are also already finding their way into commercially available applications. One such example is 'AI Dungeon', an interactive fiction and story-telling application powered by GPT-3 and its predecessor GPT-2. AI Dungeon allows users to create a variety of fictional scenarios ranging from fantasy and science-fiction interactions featuring wizards, dragons, and spaceships to more realistic settings such as interviews with famous figures both past and present (Shevlin, 2020).

It seems likely, then, that human social interactions with chatbots and other artificial systems will deepen and become more commonplace in the years to come. This possibility and its ramifications has already occasioned considerable debate in fields such as psychology and philosophy (Danaher, 2019; Perez-Osorio & Wykowska, 2020), and similar themes have long been explored in fiction, from the automaton "Olimpia" in Hoffman's 19th short story *Der Sandmann* to countless modern science fiction stories and films.

A critical point worth bearing in mind, however, is that language models and chatbots in both their current and near-term forms are still very far from exhibiting the kind of general intelligence often associated with fictional visions of future AIs and robots. In particular, such systems lack any analogue to many human cognitive capabilities such episodic or working memory, imagistic reasoning, or spatial navigation. They are non-embodied systems, without perception or bodily sensation. And perhaps most critically of all, they are non-agential, lacking any conception of short-term or long-term goals beyond returning outputs to user prompts.

Consequently, while GPT-3 can often hold a convincing conversation in short bursts and even engage in some forms of commonsense reasoning, it is also prone to a number of decidedly unhuman-like errors. It frequently struggles to keep track of key information in a conversation, such as the goals of specific named agents, and sometimes seems to lack basic knowledge of the physical properties of different objects or social norms. Perhaps most disconcertingly for its interlocutors, it is prone to contradict itself or change the topic abruptly. Such limitations frequently do make themselves felt in conversational interactions, as anyone who has spent any time conversing with a chatbot can attest. As one study puts it, "failing to adhere to social norms and glaring signs of humanity leads to decreased engagement unless balanced appropriately" (Muresan & Pohl, 2019).

While some critics (notably Marcus & Davis, 2020) have seen these limitations as casting GPT-3's abilities in a poor light, they have been acknowledged from the outset by its developers. As stressed by Brown et al. (2020) in the paper that brought GPT-3 to public attention, "GPT-3 samples still sometimes repeat themselves semantically at the document level, lose coherence over sufficiently long passages, contradict themselves, and occasionally contain non-sequitur sentences." Moreover, it should be emphasised GPT-3 was not specifically designed to be a convincing interlocutor, let alone to pass more demanding measures of humanlike conversation such as the Turing Test (Turing, 1950).

The fact that these limitations do not entirely prevent language models and chatbots from posing as at least moderately effectively conversation partners is remarkable, reflective

perhaps of the degree to which natural language encodes predictable albeit complex statistical regularities. And as systems like GPT-3 and its successors are tweaked and augmented specifically for conversational purposes, their believability seems likely to increase, albeit not to the point of capturing all the fine-grained details of human language.[2]

My goal in what follows will be to explore the significance for philosophy of mind of these distinctively *limited* systems, or as I will henceforth refer to them, *Uncanny Communicators* (UCs). While I will not give a formal definition of UCs, I will use the term deliberately broadly to refer to current and near-future natural language AIs that (i) exhibit a reasonable degree of conversational competence, capable of engaging and sustaining human users in conversations, and (ii) lack the kind of cognitive complexity, sophistication, or sensorimotor capacities to pass more demanding tests of intelligence such as stricter formulations of the Turing Test. I take it to include systems such as GPT-3 and BERT, as well as future systems that do not depart in radical ways from these architectures.

3. Does it matter if people misattribute mental states to AIs?

Thus far I have argued that it is likely that the near future of AI is likely to involve a broad and deep level of engagement between human users and chatbots in a variety of applications. If cases like Replika are anything to go by, we should expect human users in at least some contexts to treat such UCs in many of the same ways as they treat human interlocutors, attributing to them intentions, beliefs, and even emotions.

On the face of it, this amounts to a revolution in folk psychology, both because it dramatically expands the kinds of beings to which we intuitively attribute mental states, and also because many of the systems to which we will attribute mental states differ so greatly from humans in respect of their cognitive architectures. If such systems really do have beliefs and other mental states, then this should have significant ramifications for cognitive science. On the other hand, if our folk attributions are systematically in error on such a large scale, we might have grounds for concern about the consequences of such positive errors.

However, some may wish to be more dismissive about such concerns. First, one might question whether attributions of states like beliefs to AIs are really 'truth-apt' at all. While it is common in cognitive science and philosophy of mind to treat questions about the range of beings capable of possessing different mental states as being substantive and important, this sense is not universally shared, especially in computer science. As Edsger Dijkstra famously put it, "the question of whether Machines Can Think… is about as relevant as the question of whether Submarines Can Swim." A contemporary formulation of this view is given by Tom Dietterich, who has argued that these questions should not trouble the minds of AI researchers. As he puts it, "we should pursue advances in the science and technology of AI without engaging in debates about what counts as "genuine" understanding" (Dietterich, 2019).

While it might be quite reasonable for computer scientists and AI researchers

[2] Some interesting parallels should be noted with the vision of the future of AI discussed by Eric Drexler (Drexler, 2019). In short, Drexler suggests that one plausible scenario for artificial intelligence development will involve a distributed network of "comprehensive AI services" (CAIS), all individually optimized for specific tasks, rather than unified superintelligent agents. Powerful language models capable of serving as human level chatbots – but incapable of acting in the world via other means – could constitute one pillar of such a scenario.

themselves not to trouble themselves with such questions, as I have argued elsewhere (redacted for blind review) it is my firm conviction that the issue of whether UCs or other artificial systems genuinely have beliefs and other mental states is not merely a matter of semantics. On the contrary, as I will now suggest, it is a matter that is both amenable to theoretical and empirical enquiry and on which serious matters may turn.

For one, note that cognitive science has a long and quite successful track record of tackling similar cases concerning whether ascriptions of psychological states in 'non-standard' cases is justified. The project in comparative psychology of determining whether apes possess a theory of mind – an ability to attribute mental states to others – is one such example. Initially motivated by a landmark article by Premack and Woodruff (1978), several decades of research have now shed considerable light on the capacities of chimpanzees and other apes to attribute mental states to conspecifics and to humans. Along the way, the project has encountered a number of controversies and challenges, but these in turn have spurred novel experimental methods such as experience-projection tasks and ever more ingenious paradigms for teasing apart hypotheses. As matters stand, a growing consensus holds that chimpanzees possess at least some core mind-reading capacities (Krupenye, Kano, Hirata, Call, & Tomasello, 2016), though the point is not uncontested (Heyes, 2017).

Similar examples of successful inquiry into 'borderline' cases of mental state possession include questions about whether and at what age infants come to possess numerical concepts (Carey, 2009), whether corvid birds have episodic memory (Clayton, Bussey, & Dickinson, 2003), and whether racial and other biases in humans involve mere associations or full-blown unconscious beliefs (Mandelbaum, 2016). The point that I wish to stress in the present instance is that these projects have been scientifically fruitful, resulting in a better understanding both of their target questions and cognition as a whole. And while most debates about the varieties of cognition in artificial intelligence have so far remained squarely theoretical, there is reason to think that a true 'comparative psychology of artificial intelligence' is already beginning to blossom (Buckner, 2019).

Consequently, there may be good reason not simply to take folk psychological ascriptions of mental states to UCs at face value, or as an unimportant conversational extension of existing practice. Not only is there the potential for valuable science to be done, but I would also stress the danger of divergence between standards across disciplines. To the extent that it becomes normalised to attribute mental states to AIs but not to animals, for example, there is a risk that we will prioritise the interests of the former to the detriment of the latter in areas where they conflict. More broadly, one might consider that such a possibility would even threaten the 'unity of cognitive science', making it harder for psychologists and philosophers to engage in both academic and public discussion about the nature of mental states.

A second source of insouciance about the impending liberal application of psychological terms to AIs might come from a sense that such attributions are merely figurative, or not taken seriously by the users who make them. After all, we routinely anthropomorphise any number of objects that we interact with, from our cars and cellphones to individual programs running on our computers, as when we say that piece of software is confused or 'busy thinking'.

However, the sheer *earnestness* and personal concern of the reports like those of users

interacting with Replika discussed above tells against such light-weight interpretations. Moreover, there is growing empirical evidence that people really do attribute mental states to artificial systems. For example, humans viewing robots performing goal-directed behaviours such as reaching for a wine glass show similar activation in mirror neuron systems as they display when observing humans performing similar actions (Gazzola, Rizzolatti, Wicker, & Keysers, 2007; Oberman, McCleery, Ramachandran, & Pineda, 2007). Another recent study by Thellman et al. (2017) using images and verbal descriptions of human and robot behaviour asked participants to rate the agent's action in each case according to metrics of intentionality, desirability, and controllability. Results showed that subjects made similar judgments in both human and robotic cases, leading the authors to conclude that "people's intentional stance toward the robot was in this case very similar to their stance toward the human."

Similar findings have suggested a tendency for humans to *empathise* with robots. Studies by Rosenthal-von der Pütten et al. (Rosenthal-von der Pütten, Krämer, Hoffmann, Sobieraj, & Eimler, 2013) and Suzuki et al. (Suzuki, Galli, Ikeda, Itakura, & Kitazaki, 2015) found similar electroencephalographic and arousal responses from participants observing both human and robots in apparent pain. Evidence from humans who interact with robots on a daily basis in life-and-death situations bears this out, most famously perhaps in the responses of armed service personnel to 'Packbots', drones used in the field for detecting and dismantling improvised explosive devices. Philosopher David Gunkel, summarizing reports of such interactions, notes the practice of soldier "giving [Packbots] names, awarding them battlefield promotions, risking their own lives to protect that of the robot, and even mourning their death" (Gunkel, 2018).

Such examples do not, of course, show that our intuitive attributions of mentality to artificial systems are well-grounded, but I do take them to suggest that in at least some cases our interactions with robots lead us to genuinely impute mental states to them in a way that can scientifically and behaviourally distinguished from merely figurative anthropomorphism. Putting matters simply, most people do *not* in practice treat the issue of whether robots can have mental states as analogous to the question of whether a submarine can swim, but either exhibit or fail to exhibit patterns of response to artificial systems depending on whether they are treating them as agents or as mere machines. In turn, this makes more urgent and intelligible the project for cognitive scientists of determining whether and in which cases such responses are justified by the cognitive capacities of the system in question.

A third reasons for the dismissal of the significance of the trend towards mentalising AIs and UCs might be grounded in a sense that these attributions might not matter very much, even if they are intended literally and are scientifically- and philosophically speaking, unjustified. Against this, I would suggest that a number of social and ethical issues really do seem to hinge on such determinations. Some of these are of a more sociological nature, such as the concerns about alienation, isolation, and neglect that might arise from an individual's choosing to spend time with actually insensate robot companions rather than family, friends, or dependents (however, see Danaher, 2019).

Another set of concerns stem from the likely possibility that we will assign moral or even political rights to AIs, as highlighted by the decision of Saudi Arabia to grant citizenship to the robot 'Sophia'. While such cases are for now largely a matter of public spectacle, as

people form deeper attachments to robots and other AI systems it is likely only a matter of time before moral and even legal arguments are made for affording them rights.

Given the history of human disregard for animals and even other humans, such a move may seem well-intentioned and relatively benign. It might also be justified on the basis of inductive risk arguments and a 'precautionary principle' (Birch, 2017; Douglas, 2000) in recognition of the fact that the risks of positive errors in attributing moral status to machines are likely to be lower than concomitant negative errors. However, great care must be taken here, for several reasons. For example, if we attribute moral status to machines without good psychological basis for doing so, this may be reflected in our social policies (such as allocations of funds) in a way that leads to us to neglect opportunities for improving the well-being of humans and animals with a much stronger philosophical and scientific claim to genuine moral status. Perhaps more subtly, there is a risk than in misattributing moral status to machines that do not deserve it, we will be more likely to miss indicators of machines that *do* deserve such consideration. Simply leaving matters to intuition, for example, it is likely that even quite stupid but highly social AIs with human-like characteristics would be given moral status, while smarter but less anthropomorphic AI with genuine capacity for suffering or undergoing harm would be missed (Shevlin, 2021; Tomasik, 2014).

Concerns about the consequence of over-attribution of mentality to artificial systems also apply to our attributions of intentional states like beliefs, desires, and intentions. most notably concerning whether we are ever justified in treating machines as epistemic or moral agents. For example, one forthcoming study (Kneer, 2021) used a scenario involving a robotic receptionist who falsely tells a visitor that the manager is absent, when in fact she is present but merely wishes not to be disturbed. The study found that participants attributed deceptive motives to the robotic receptionists at almost the same frequency as their human equivalents.

Interestingly, a scenario was also included in which (unbeknown to the receptionist) the manager was indeed absent, hence allowing the study to distinguish between two kinds of deception, one involving conveying information that is both believed to be false and is indeed false, and the involving conveying information that is believed to be false but is in fact true. This is important insofar as it ensures that subjects responses were tracking the assumed *intention* of the agent rather than merely whether it gives true or false information. Critically, subjects overwhelmingly imputed deception in both cases, suggesting that their judgments were indeed driven by ascriptions of intention specifically.

While this study involved verbally-presented scenarios, its implications are potentially quite significant. If people are indeed willing to attribute deceptive intentions to an artificial system, it raises the concern that human actors may similarly *fail* to be held to account for deceptions employed by their AI helpers. As noted by the author of the paper, "[i]f robots are judged as capable of lying, and are attributed… blame for this behavior, human agents who instrumentalize them in a wide range of domains from deceptive marketing to political smear-campaigns might be judged less blameworthy than they actually are."

A similar but even more astonishing study found that participants in a fictional scenario involving an AI whose actions resulting in poisoning of humans were apt to *morally blame* the artificial system for its actions (Kneer & Stuart, 2021). More specifically, the study found that while humans were disinclined to apportion blame to an 'unsophisticated' AI, they readily assigned responsibility to 'sophisticated' and 'semi-sophisticated' systems.

It is worth stressing that the 'semi-sophisticated' system in particular was not some fanciful far-future artificial general intelligence, but was simply stipulated to possess a small repertoire of case-relevant concepts such as POLLUTON and HUMAN LIFE, and as being "capable of language-based interaction and [making] hypotheses about human mental states and tests them against observations (i.e., it has "theory of mind")." These capacities are not radically removed from the capabilities of some emerging artificial systems, such as the impressive predictive capacities of Google's "ToM NET" (Rabinowitz, Perbet, Song, Zhang, & Botvinick, 2018).

Again, the tendency of humans to hold artificial agents accountable raises broader ethical concerns about potential misuse by humans eager to avoid responsibility. This potential implication was specifically tested for in the aforementioned study, and it was found that companies employing *less* sophisticated AIs were held more responsible for their own actions, suggesting that some 'share' of the blame was being attributed to AIs in the scenarios involving more sophisticated systems.

The relevance of these studies for the present project is, I hope, fairly clear: people are willing to attribute a range of relatively sophisticated cognitive states to artificial systems, and these in turn influence the "reactive attitudes" (Strawson, 2008) they adopt towards them, with knock-on effects for their treatment of involved human agents. To the extent that these attributions are based on ignorance or falsehoods of the genuine psychological capacities of the system in question, their moral responses may in turn be misguided (however, see Coeckelbergh, 2009, for a dissenting view). Of course, it is also possible – especially when considering medium- and long-term AIs – that these ascriptions of moral culpability are justified. Either way, matters of no small significance may hinge on the question.

4. Uncanny believers?

Thus far, I have made several two main sets of claims. The first is that we have good reason to believe that the emerging domain of social chatbots will expand and improve in such a way that an ever larger number of people engage with UCs in ways that involve attributing to them beliefs, desires, feelings, and other mental states. The second is that such attributions *matter*: they are not merely semantic issues, nor are they (in some cases at least) merely figurative, and a number of issues of social import hang upon whether these attributions are correct. Now, in this final part of the paper, I will argue that we have multiple good grounds for thinking such attributions are and will be made erroneously by the lights of contemporary cognitive science and philosophy, at least in regard to belief specifically.

I have chosen to focus on beliefs rather than perhaps more obviously ethically weighty states such as agency or consciousness for several reasons. First is the simple consideration that beliefs are perhaps at least slightly more tractable and less immersed in metaphysical and normative debates than mental capacities such as consciousness and agency. Artificial consciousness, needless to say, is a vast and controversial topic well beyond the scope of this paper, and there is considerable doubt among some philosophers as to whether any artificial system could be conscious (Godfrey-Smith, 2016). This in turn creates a difficulty when we wish to consider whether an artificial system could have undergo mental states that are closely linked to consciousness, such as pain or emotion.

A second reason for focusing on beliefs is that they are a widely held to be closely connected to a wide range of other mental states of interest, including intention, agency, and

theory of mind, such that if we had good reason not to attribute beliefs to a system, it may also not make sense to attribute these other states. It would be beyond the scope of the present paper to make any strong claims about the precise relations between these states, but it is fair to say that a system incapable of forming any beliefs would not qualify as having intentions or agency by the lights of most contemporary accounts of these phenomena.

Finally, to a greater extent than other mental states, beliefs and other propositional attitudes seem to exhibit very close connections with behaviour, especially verbal behaviour. Thus while we can arguably make sense of the idea of a 'philosophical zombie', a functional duplicate of a human that was behaviourally indistinguishable from us but lacked qualia or phenomenal consciousness, it is much harder to make sense of the idea that such a being would lack beliefs all together, even though they occur unconsciously.[3] This is arguably relevant for UCs insofar as they seem relatively better candidates for satisfying some of these more behavioural criteria than conditions that might be necessary for other mental states.

With these considerations in mind, the question before us, then, is whether UCs have any remotely reasonable claim to possess beliefs. I will be using this term in what I take to be a relatively neutral sense to capture the everyday ascriptions of beliefs we make to one another, but with a particular emphasis on what are sometimes called *occurrent explicit* beliefs. Speaking somewhat loosely, imagine that someone prompts me to answer the question as to who I think will win the next American election; at that moment, it seems that I form or access some form of mental representation that corresponds to the answer I will give in conversation.[4]

Hence imagine that we ask a chatbot the question "Where do penguins live?" with the answer "They live predominantly in Antarctica". The question we are concerned with, then, is whether we might *correctly* report this output by saying of that chatbot that it believes that penguins live in Antarctica. To answer this question, we might appeal to two broad 'families' of views about beliefs that dominate much of the current literature, and which following Schwitzgebel (2013) we might term *deep* and *superficial*.

In short, deep views of belief take questions about whether an individual possesses a belief to boil down to questions about whether they stand in an appropriate relation to a representation with the right kind of content and (perhaps) poised to play the right kind of cognitive role (e.g., Carruthers, 2006; Dretske, 1991; Fodor, 1987; Millikan, 1984; Quilty-Dunn & Mandelbaum, 2018). At its core is the view that beliefs and other folk psychological states are constituted by states of a system's internal structure that interact mechanistically in ways that produce the generalisations identified by folk psychology. As Fodor (1981) puts it, "to suppose two people share a belief is to suppose them to be ultimately in some structurally similar internal condition, e.g., for them to have the same words of Mentalese written in the

---

[3] We might, of course, wish to allow that such a philosophical zombie would have beliefs with substantially different content from ours, especially in regard to its own mental states (Chalmers, 1996). Likewise, we might wish to assign a significant role to interactions with the external environment in grounding the specific content of our beliefs (Putnam, 2011).

[4] Other beliefs – sometimes called dispositional beliefs – might not be activated in thought at a given moment but nonetheless serve in psychological explanations. For example, if I asked why Jane seems unconcerned about volcanic eruptions interrupting her daily business, I might reasonably say that she believes that there are no active volcanos in the vicinity of her home in Greenwich Village, even if that particular thought has not 'crossed her mind', so to speak, in recent memory.

functionally relevant places in their brains."

Typically (but not always) deep accounts of beliefs appeal to a broadly computational theory of the mind according to which thinking, believing, desiring, and so on states constitutively involve computational operations across representations. One of the leading virtues of such accounts is that can at least in principle give plausible mechanistic explanations for many of the phenomena characteristic of human thought, such as systematicity and productivity, and the fact that we can undergo various mental states with shared contents. Thus the reason I can think "John is the bathroom" and I can also think "I want John to leave the bathroom" is because I can appropriately combine in various ways mental representations corresponding to myself, John, bathrooms, and so on.

By contrast, superficial accounts of belief possession do not make successful belief ascriptions so directly contingent on specific facts about the state of an individual's mental architecture at a given time. Rather, they take an individual's possession of a given belief to be contingent on *surface-level* properties, that is, properties manifest in its broader behaviour and dispositions. To give a simple example, a personality trait such as 'being witty' is likely be superficial in this sense: there is no brain area, set of representations, or particular patterns of computation constitute of being witty; rather, to be witty is (in some complex way) to display particular kinds of verbal and physical behaviour.

For some superficial theorists (Davidson, 2001; Daniel Clement Dennett, 1989; Ryle, 2009) the relevant surface-level properties for belief attribution are restricted specifically to publicly observable behaviour. For others (notably Schwitzgebel, 2002, 2013) these surface-level properties may also include *phenomenal* and *cognitive* properties. While these may not be directly observable in the same way as public behaviour, these still 'count' as superficial insofar as they are a varied, complex, and dynamic set of properties that can be variously instantiated, as opposed to deep features whose existence is tied to particular well-specifiable facts about the role played by particular sorts of representations within cognitive architectures. Superficial accounts of belief for their part have a number of attractive features, such as their ability to give intuitively correct accounts of belief in deviant cases (Schwitzgebel, 2013), as well as avoiding commitment to the idea that we have a particular number of beliefs with determinate content 'stored in our brains' at any one time. Moreover, as Dennett argues, superficial accounts capture key features of folk psychology, such as the fact that "[f]olk psychology is abstract. It requires conceptual knowledge but not concrete causal knowledge."

Needless to say, within the two families are a vast number of views that make quite different commitments, and a thorough exploration of what each such theory might have to say about beliefs in AIs like UCs would be well beyond the scope of this paper. However, I think we can begin to make some useful generalisations about their broad applicability to the question.

Consider first deep theories of belief. While historically many deep theories have been sympathetic to the idea that artificial systems could have precisely the same kinds of states as humans, in the case of UCs in particular few deep theorists are likely to assent to attributions of belief being well grounded. In short, this is because the particular computational role played by representations within contemporary systems are significantly unlike similar representations in the human mind.

We can spell this out a little more by pointing to three major areas of difference. The

first is that UCs' states are very unlikely to possess content of the kind invoked in the folk psychological attributions we might naively be inclined to make to them. Unlike contemporary computer vision systems trained on natural images, language models and chatbots have never seen or directly interacted with the objects they describe in their outputs. Words, for language models, exist as objects with statistical relation to other words, rather than as corresponding to anything in the world. For deep theorists such as Dretske and Fodor committed to causal views of content, this would tell against the idea that activations in language models are strictly speaking *about* anything at all. Likewise, the fact that UCs are artefacts severed from any evolutionary history would mean that teleosemanticists such as Millikan would reject imputations of content to their processes in any robust *original* sense. This is not to deny, of course, that we can interpret their outputs as meaningful, but this a fact about us rather than the world.

Second, UCs are likely to lack the same kinds of representational structure as those that deep theorists take to be constitutive of human belief production and updating. Again, details here are complex and will vary across different models and theories, but insofar as models are trained exclusively on large databases of text, their atomic unit of thoughts are words, sentences, and strings of text, rather than representations able to be expressed variously in natural language. There is nothing in large language models that identifies the inputs "2" and "two" as picking out the same concept, let alone "dos" or "deux", or "the smallest prime number", for example; they are simply treated as different possible inputs and outputs, albeit ones that may in some cases have similar statistical properties.

Finally, UCs lack many of the high-level psychofunctional characteristics of human belief. To give one example, humans typically hold a wide variety of beliefs of varying strengths that are relatively independent of what was said in immediately preceding conversation: we have psychologically central beliefs about our identity, our past, science, the history of the world, and our values, as well more peripheral beliefs about the weather and what we are going to have for dinner. By contrast, a characteristic feature of chatbots is their conversational malleability: the views they adopt and reject are largely a function of the conversation as it has unfolded thus far. The fact that humans have such a 'web of belief' (Quine & Ullian, 1978) is arguably a core feature of what we mean by belief in the first place, one that UCs almost entirely lack. Of course, there are many other important psychofunctional aspects of human belief that may be more or less central to the phenomenon that UCs all together lack. The psychofunctional theory of belief sketched by Mandelbaum and Quilty-Dunn (2018) for example, "identifies beliefs as relations to mental representations, with the relations characterized by the psychological generalizations that hold over belief. These generalizations include that beliefs are acquired ballistically and automatically, put subjects into a negatively valenced motivational state when encountering disconfirming evidence, are changed in ways that will assuage that state, will increase in strength over time if left alone, will increase in strength even more if repeatedly tokened, and will increase in accessibility the more they are activated." Needless to say, most of these are unlikely to hold true of the functional dynamics of Large Language Models.

Let us turn, then, to superficial accounts. Again, of course, we are dealing with a wide array of theories with different commitments, but a few broad lessons can be drawn. On the face of it, superficial accounts seem at least in principle more amenable to the literal truth of belief ascriptions to UCs, insofar as chatbots at least seem to exhibit many of the same broad

forms of verbal behaviour as humans. To give an oversimplified example, imagine that a superficial theorist identified the belief that penguins live in Antarctica with a set of verbal dispositions including, for example, the disposition to respond "Antarctica" to the question "where do penguins live", to name penguins when prompted to list animals that live in Antarctica, and so on. By this (admittedly crude) metric, a chatbot might indeed qualify as a 'true believer', at least as far as penguins are concerned.

This may too fast, however. After all, most superficial theorists place a considerable weight on non-verbal dispositions as well as verbal dispositions. To continue the toy example above, this might mean that to believe that penguins live in Antarctica would also involve dispositions like going to the 'Antarctic Animals' section of a zoo when asked to locate the penguins. Clearly, no UC will be able to satisfy dispositions like these.

One might object that such nonverbal dispositions can hardly be *necessary* parts of the cluster of dispositions required for ascribing beliefs, not least because many humans may fail to exhibit them. Someone who believed that the zoo 'animals' were in fact animatronic replicas might not go to the Antarctic section of the zoo when asked to find penguins, for example. More drastically, one might consider humans whose non-verbal behavioural repertoire is limited by virtue of disability. A person with 'locked-in syndrome', for example, might be unable to engage in almost all forms of non-verbal behaviour by virtue of paralysis, yet this would hardly prevent us from ascribing beliefs to them on the basis of their verbal utterances made via assistive technologies.[5] Such cases might lead a superficial theorist to conclude that non-verbal behaviour was inessential for the possession of beliefs, and thus could potentially be ascribed to UCs. If one thinks of a superintelligent artificial system lacking sensorimotor capacities but easily capable of passing the Turing Test (perhaps akin to *2001*'s 'HAL'), the mere fact that it was incapable of moving through the world would probably not deter us from attributing beliefs to it.

Still, there is likely to be considerable variation in the willingness of different superficial theorists to attribute beliefs to UCs. Some superficial approaches, notably Schwitzgebel's phenomenal-dispositional account (2002, 2013), would place equal emphasis both on behavioural dispositions and on *cognitive* and *phenomenal* dispositions. To give a simple example, on Schwitzgebel's view, to believe that there is beer in the fridge is – in addition to being disposed to say things like 'there is beer in the fridge' – to also "feel surprise should one go to the fridge and find no beer" and "to draw conclusions entailed by the proposition that there is beer in the fridge (e.g., that there is something in the fridge, that there is beer in the house), and so forth" (Schwitzgebel, 2002: 251). This is bad news for UCs: in addition to presumably lacking phenomenal consciousness all together (and being incapable of *feeling* surprise or anything else), their cognitive dispositions (if they can be described as such) are very different from ours. This can be readily observed in the frequent patterns of non-sequiturs, confusions, and misunderstandings demonstrated by even our best current chatbots.

Perhaps the most liberal form of superficial theory is Dennettian 'interpretationism'.

---

[5] One complication with this example comes from complex issues concerning the attributions of dispositions. A patient with locked-in syndrome would not be able to reach for a glass of water if they were thirsty, but should we conclude from this that they lacked the disposition to do so, or rather that they were simply unable to actualize their dispositions due to their disability?

This is a nuanced and complex view, and its exact details will not be spelled out here, but in short it treats belief ascriptions as justified to the extent that by adopting the "intentional stance" towards a system, we can gain novel insights into its behaviour that could not be had as easily or as cheaply elsewhere. To give one of Dennett's examples (1989), imagine that we overhear the phone ring in Mrs Gardner's kitchen. he telephone rings in Mrs Gardner's kitchen. She answers, and says "Oh, hello dear. You 're coming home early? Within the hour? And bringing the boss to dinner? Pick up a bottle of wine on the way home, then, and drive carefully". From that simple observation, by adopting the intentional stance we can conclude that "a large metallic vehicle with rubber tyres will come to a stop in the drive within one hour, disgorging two human beings one of whom will be holding a paper bag containing a bottle containing an alcoholic fluid." While an observer incapable of adopting the intentional stance might come to the same conclusion (perhaps via some dizzyingly complicated prediction concerning the likely states of different fundamental particles in an hour's time), to do so would require vastly more data and computation than the simple application of folk psychology. It is in virtue of the predictive power afforded by the adoption of the intentional stance in this case, then, that we can say that the actors in this example *really do* have mental states of the relevant kind.

We must be a little careful here, for Dennett's exact view on the metaphysical status of beliefs is notoriously subtle (Dennett, 1991). On the face of it, by making the truth of belief-ascriptions purely a matter of their utility, it may seem that he is endorsing a kind of anti-realism about beliefs as entities in their own right. However, he is at pains to deny this. While he rejects what he calls Fodor's "industrial strength Realism", he insists that the success of the intentional stance as an explanatory and predictive strategy stems from the fact that it identifies "real patterns" in the world, patterns which would be missed were we to true to understand the world purely in physical mechanistic terms.

What is clear is that Dennettian interpretationism is avowedly liberal in its willingness to countenance beliefs in a wide range of systems. As Dennett (1989) puts it, "[m]y view embraces the broadest liberalism, gladly paying the price of a few recalcitrant intuitions for the generality gained." Insofar as people can and do successfully attribute beliefs to artificial systems and thereby gain new insights into their likely behaviour, then, such attributions will be *ipso facto* justified.

Should we conclude from this that Dennettian interpretationism may constitute a way of reconciling philosophical and folk psychological views of beliefs in artificial systems like UCs? That would be too quick, I think. Even if it is possible for UCs to have beliefs, it does not follow that the beliefs they really possess on a Dennettian view are *the same beliefs* that we naively ascribe to them. Users of chatbots like Replika are frequently confounded when the system fails to remember facts about them they have previously told it, or changes its apparent personality, or otherwise displays verbal behaviour at odds from that which they had been expecting. This surprise and consternation suggests a *failure* of the intentional stance in such cases.

To make the point clearer with an example, imagine that we were to encounter a human who claimed to believe that the Apollo Moon Landings were faked. When asked further about the matter, however, they sometimes seemed to change their mind, or to alter the subject of the conversation. In such a case, we might reasonably hesitate to ascribe them the belief they had initially seemed to profess. Instead, we might attempt to understand the

utterance as a joke, or a performance, or some kind of confusion. Given the tendency of even the best chatbots to make similar sorts of errors, we might conclude that instead of attributing to them robust stable beliefs as we might be initially inclined to do, we should attribute much vaguer or more transitory attitudes. To the extent, then, that our interactions with UCs continue to vex and surprise us, then, we have good reason to think that even on liberal views like Dennett's, our initial temptation to attribute specific human-like beliefs to them was in at least some cases a mistake. A gap would still remain, then, between the judgments of our philosophical and scientific theories of belief and folk psychological practice in our engagement with chatbots and UCs.

5. Conclusion: closing the gap between folk psychology and cognitive science

This paper has made three main claims. The first is that we are likely to see a rise in their depth and breadth of popular engagement with social chatbots, and that such engagement will be (and indeed, already is) marked by the frequent attribution of mental states to the systems in question. The second is that these attributions should not be dismissed as merely conventional or figurative, but rather as implicitly committing to real claims about the nature of these systems, claims that in at least some cases will have socially and ethically significant consequences. Finally, I argued that in the case of belief at least, according to the lights of the dominant theories in philosophy and cognitive science these claims are strictly speaking *false*, or at least liable to be inaccurate and misleading in many cases.

If these claims are true, I take it that an interesting problem faces us concerning how we might bridge this growing gap between what the folk intuitively believe about UCs' psychological states and the considered judgments of scientists and philosophers about the matter. What should be done? On the one hand, we might attempt to *re-educate* the folk, trying to disabuse them of the idea that the systems in question really have the kinds of states they attribute to them. Such an endeavour, however, seems at least *prima facie* unlikely to succeed, especially given the clear commercial incentives that will exist for developing chatbots that exploit our anthropomorphizing tendencies to maximise user engagement.

Another possibility would be to simply accept that the folk have 'moved on', so to speak, from the concepts of belief that have dominated cognitive science and philosophy thus far, and that have in effect reappropriated the concept for usage in a new way. To avoid confusion, we might even decide to alter our own technical vocabulary, talking no longer of beliefs simpliciter, but perhaps *Cognitive Beliefs*, or some other phraseology designed to make explicit the mismatch between the terms used in cognitive science and their folk psychological equivalents.

Whichever option is adopted, academics should certainly try to ensure that policymakers and stakeholders in legal debates are not overly influenced by folk psychological practice in domains such as corporate responsibility for artificial systems. More broadly, the rapid expansion or extension of folk psychological terms to UCs and other systems with architectures radically different humans is a notable development in our society's history, and one likely to have strange and unpredictable consequences. Some of these may be benign, while others require more direct action to limit their negative consequences. Whatever the case may be, it is likely to create strange, unfamiliar, and indeed uncanny situations requiring further engagement from philosophers.

REFERENCES

Askell, A., Clark, J., Brundage, M., Hernandez, D., Lansky, D., Luan, D., … Wu, J. (2019). Better Language Models and Their Implications.

Balch, O. (2020). AI and me: friendship chatbots are on the rise, but is there a gendered design flaw? Retrieved February 13, 2021, from The Guardian website: https://www.theguardian.com/careers/2020/may/07/ai-and-me-friendship-chatbots-are-on-the-rise-but-is-there-a-gendered-design-flaw

Birch, J. (2017). Animal sentience and the precautionary principle. *Animal Sentience: An Interdisciplinary Journal on Animal Feeling*, *2*(16), 1–16. Retrieved from http://animalstudiesrepository.org/animsent/guidelines.html

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *ArXiv*.

Buckner, C. (2019, May). The Comparative Psychology of Artificial Intelligences. Retrieved August 5, 2019, from http://philsci-archive.pitt.edu/16034/

Carey, S. (2009). The Origin of Concepts. In *The Origin of Concepts*. https://doi.org/10.1093/acprof:oso/9780195367638.001.0001

Carruthers, P. (2006). *The architecture of the mind*. Oxford University Press.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. https://doi.org/10.1002/(sici)1097-461x(1998)66:1<107::aid-qua9>3.0.co;2-z

Clayton, N. S., Bussey, T. J., & Dickinson, A. (2003). Can animals recall the past and plan for the future? *Nature Reviews Neuroscience*, *4*(8), 685–691. https://doi.org/10.1038/nrn1180

Coeckelbergh, M. (2009). Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception, and performance of artificial agents. *AI and Society*. https://doi.org/10.1007/s00146-009-0208-3

Danaher. (2019). The Philosophical Case for Robot Friendship. *Journal of Posthuman Studies*, *3*(1), 5. https://doi.org/10.5325/jpoststud.3.1.0005

Davidson, D. (2001). *Inquiries Into Truth and Interpretation: Philosophical Essays Volume 2* (Vol. 2). Oxford University Press.

Dennett, Daniel C. (1991). Real Patterns. *The Journal of Philosophy*. https://doi.org/10.2307/2027085

Dennett, Daniel Clement. (1989). *The intentional stance*. MIT press.

Dietterich, T. G. (2019). What does it mean for a machine to “understand”? Retrieved February 14, 2021, from Medium website: https://medium.com/@tdietterich/what-does-it-mean-for-a-machine-to-understand-555485f3ad40

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*. https://doi.org/10.1086/392855

Dretske, F. I. (1991). *Explaining behavior: Reasons in a world of causes*. MIT press.

Drexler, K. E. (2019). *Reframing Superintelligence: Comprehensive AI Services as General Intelligence*. Retrieved from https://www.fhi.ox.ac.uk/reframing/

Epstein, R. (2007). From Russia, with Love. *Scientific American Mind*, *18*(5), 16–17. https://doi.org/10.1038/scientificamericanmind1007-16

Fodor, J. (1981). Representations: Philosophical Essays on the Foundations of Cognitive Science. Retrieved May 29, 2019, from https://philpapers.org/rec/FODRPE

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind* (Vol. 2). MIT press.

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2007.02.003

Godfrey-Smith, P. (2016). Mind, matter, and metabolism. *Journal of Philosophy*.

https://doi.org/10.5840/jphil20161131034

Gunkel, D. J. (2018). The other question: can and should robots have rights? *Ethics and Information Technology*. https://doi.org/10.1007/s10676-017-9442-4

He, P., Liu, X., Gao, J., & Chen, W. (2021). Microsoft DeBERTa surpasses human performance on the SuperGLUE benchmark. Retrieved February 13, 2021, from Microsoft Research Blog website: https://www.microsoft.com/en-us/research/blog/microsoft-deberta-surpasses-human-performance-on-the-superglue-benchmark/

Heyes, C. (2017). Apes Submentalise. *Trends in Cognitive Sciences*, *21*(1), 1–2. https://doi.org/10.1016/j.tics.2016.11.006

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., … Amodei, D. (2020). Scaling laws for neural language models. *ArXiv*.

Kneer, M. (2021). Can a robot lie? *Researchgate*. https://doi.org/10.13140/RG.2.2.11737.75366

Kneer, M., & Stuart, M. T. (2021). *Playing the Blame Game with Robots*. https://doi.org/10.1145/3434074.3447202

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114. https://doi.org/10.1126/science.aaf8110

Luka, I. (2017). Three Myths About Replika. Retrieved February 13, 2021, from https://medium.com/@replika/three-myths-about-replika-7717c2d2237

Mandelbaum, E. (2016). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Noûs*, *50*(3), 629–658. https://doi.org/10.1111/nous.12089

Marcus, G., & Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. Retrieved February 13, 2021, from MIT Technology Review website: https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. MIT press.

Muresan, A., & Pohl, H. (2019). Chats with bots: Balancing imitation and engagement. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290607.3313084

Nuruzzaman, M., & Hussain, O. K. (2018). A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks. *Proceedings - 2018 IEEE 15th International Conference on e-Business Engineering, ICEBE 2018*. https://doi.org/10.1109/ICEBE.2018.00019

Oberman, L. M., McCleery, J. P., Ramachandran, V. S., & Pineda, J. A. (2007). EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2006.02.024

Pardes, A. (2018). The Emotional Chatbots Are Here To Probe Our Feelings. *Wired*.

Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2017). Deeper attention to abusive user content moderation. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. https://doi.org/10.18653/v1/d17-1117

Perez-Osorio, J., & Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*. https://doi.org/10.1080/09515089.2019.1688778

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Putnam, H. (2011). Meaning and reference. In *The Pragmatism Reader: From Peirce through the Present*. https://doi.org/10.5840/swjphil198011111

Quilty-Dunn, J., & Mandelbaum, E. (2018). Against dispositionalism: belief in cognitive science. *Philosophical Studies*, *175*(9), 2353–2372. https://doi.org/10.1007/s11098-017-0962-x

Quine, W. V. O., & Ullian, J. S. (1978). *The web of belief* (Vol. 2). Random house New York.

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., & Botvinick, M. (2018). Machine Theory of mind. *35th International Conference on Machine Learning, ICML 2018*, *10*, 6723–6738. Retrieved from https://arxiv.org/abs/1802.07740

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An Experimental Study on Emotional Reactions Towards a Robot. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-012-0173-8

Ryle, G. (2009). *The concept of mind*. Routledge.

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Nous*. https://doi.org/10.1111/1468-0068.00370

Schwitzgebel, E. (2013). A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box. In *New Essays on Belief*. https://doi.org/10.1057/9781137026521_5

Shevlin, H. (2020). A Digital Remix of Humanity. Retrieved February 27, 2021, from https://dailynous.com/2020/07/30/philosophers-gpt-3/#shevlin

Shevlin, H. (2021). How Could We Know When a Robot was a Moral Patient? *Cambridge Quarterly of Healthcare Ethics*.

Shevlin, H., & Halina, M. (2019). Apply rich psychological terms in AI with care. *Nature Machine Intelligence*, *1*(4), 165–167. https://doi.org/10.1038/s42256-019-0039-y

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My Chatbot Companion - a Study of Human-Chatbot Relationships. *International Journal of Human Computer Studies*. https://doi.org/10.1016/j.ijhcs.2021.102601

Strawson, P. F. (2008). Freedom and resentment and other essays. In *Freedom and Resentment and Other Essays*. https://doi.org/10.4324/9780203882566

Suzuki, Y., Galli, L., Ikeda, A., Itakura, S., & Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports*. https://doi.org/10.1038/srep15924

Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2017.01962

Tomasik, B. (2014). *Do Artificial Reinforcement-Learning Agents Matter Morally?*

Turing, A. M. (1950). Computer Machinery and Intelligence. *Mind*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., … Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *ArXiv*.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation.*

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … Dean, J. (2016). Google's NMT. *ArXiv E-Prints*.