


LLMS LACK A THEORY OF MIND AND SO CAN'T PERFORM SPEECH ACTS—A CAUSAL ARGUMENT

Justin Tiehen *

University of Puget Sound, Philosophy Department

*jtiehen@pugetsound.edu

Abstract I advance a causal argument for the conclusion that large language models (LLMs) lack Theory of Mind and so can't perform speech acts. The argument is causal in that the animating idea is that LLMs are unable to learn or understand causal relations, a claim that I support by drawing on the views of Judea Pearl. I argue that if LLMs have this sort of causal problem, it follows that they cannot possess Theory of Mind, given the further premise that Theory of Mind should be understood as a causal theory. I then argue that if LLMs lack Theory of Mind, they cannot perform speech acts, given the further premise that speech acts require reflexive-communicative intentions and so require Theory of Mind. As a consequence of the result that LLMs cannot perform speech acts, I conclude that their utterances lack speaker meaning. Something that sets my case apart from other defenses of speech act skepticism for LLMs is that my key claim is not that LLMs cannot form intentions at all, but rather that they cannot form intentions (or other mental states) with a certain type of content—content that is about mental states. After setting out my argument, I consider an objection to my view based on the observation that LLMs perform fairly well on Theory of Mind benchmarks. I reply to this objection by proposing that LLMs rely on a different type of mechanism than human beings do to achieve this performance. I support this proposal by pointing to a form of behavior that LLMs do not exhibit but that we should expect them to exhibit if they possessed Theory of Mind.

Keywords: Theory of Mind, Causal Inference, Speech Acts, LLMs, Judea Pearl, Interventions

1. Introduction

In this paper, I advance a causal argument for the conclusion that large language models (LLMs) like OpenAI's ChatGPT lack a Theory of Mind (ToM) and so can't perform speech acts. Speech act skepticism for LLMs has been defended by other authors, including Patrick Butlin and Emanuel Viebahn (2023). However, whereas these authors argue that LLMs cannot perform speech acts on the basis that they supposedly cannot form intentions, I take a different line. I take no stand on whether LLMs can form intentions as such, but instead argue they cannot form intentions (or other mental states) *with a certain type of content*—namely, content that is about mental states. This follows from their lack of ToM. Given

that there is controversy over just what is needed to have an intention, with some dispositional accounts potentially allowing that LLMs have what it takes (because they have the needed dispositions), it is an advantage of my approach compared to alternative forms of speech act skepticism about LLMs that I can avoid this controversy. My view is controversial in other respects, however. There are now a number of studies showing that LLMs do fairly well on standard ToM benchmarks. It will be a burden of my argument to explain these results away, to show that they are consistent with my view.

My argument is causal in that the animating idea is that LLMs are unable to learn or understand causal relations in general. I make the case for this claim in Section 2 by drawing on the views of Judea Pearl. I stand by this case, but take my main novel contribution in the paper to be what follows after Section 2, where I draw out the implications of the causal point for other matters. In Section 3, I argue that if I am right that LLMs have a causal problem, it follows that they cannot possess ToM, given the further premise that ToM should be understood as a causal theory. In Section 4, I argue that if LLMs lack ToM, it further follows that they cannot perform speech acts because doing so requires Gricean reflexive-communicative intentions, and so requires ToM. I also argue that it follows that the utterances of LLMs lack what is known as speaker meaning, which vindicates one way of spelling out the idea that they are mere “stochastic parrots.” In Section 5, I consider the objection to my view that I have already alluded to, according to which LLMs’ fairly successful performance at false-belief tasks and other ToM benchmarks refutes the claim that they lack ToM. In Section 6, I reply to this objection by arguing that there is good reason to think that LLM performance on such benchmarks is not due to their possessing the same type of cognitive mechanism that human beings use in performing ToM tasks, and more specifically, that it is not due to LLMs’ possessing a causal theory of other minds.

2. The Causal Limits of LLMs

Judea Pearl has developed a highly influential formal framework for understanding causal inference (Pearl 2022), a framework that has shaped work in philosophy (Spirtes et al. 2000; Woodward 2003) and psychology (e.g., Sloman 2005; Gopnik 2022) in addition to his own field of computer science. Pearl has claimed that it is a consequence of his approach that standard *deep learning models* in AI are incapable of learning or understanding causal relations despite the impressive advances such models have made over the last decade in domains like machine vision and natural language processing (Pearl and Mackenzie 2018, Chapter 10). Because LLMs are a type of deep learning model, Pearl’s claim about a causal limitation extends to them, as he has explicitly noted (Mackenzie 2023). Pearl’s view about deep learning models does not turn on their architectural details, on matters like what self-attention amounts to, for instance. Instead, it is based on in-principle *learnability* considerations. He says that no intelligent system, whether human or machine, can learn causal relations purely on the basis of the sort of non-causal statistical data that standard deep learning models are trained on. To develop this thought, in this section I review Pearl’s formal framework for understanding causal inference, his *causal ladder* (Pearl 2022; Pearl and Mackenzie 2018).

The ladder is a 3-level hierarchy that consists of increasingly richer formal languages used for describing the world. At the bottom is the Associational Level, which has vocabulary for expressing non-causal statistical relations, including especially terms for conditional probabilities like $P(y|x)$. The Associational Level is non-causal in that it does not have terms adequate for capturing causal relations. Still, it is included in the ladder because the statistical relations it can capture are often evidence that confirms or disconfirms causal hypotheses.

The next rung in the ladder is the *Interventional Level*, whose vocabulary includes everything from the Associational Level plus Pearl’s *do-calculus* (Pearl 2022, Chapter 3). This calculus features the *do-operator*, which figures in expressions like $P(y|do(x))$, representing the probability that variable Y will take on the value y given a *causal intervention* that sets the value of X to x . I will say more about this momentarily, but first let me round things off by adding to the mix the top of the ladder, the *Counterfactual Level*, which includes everything from the Interventional Level plus terminology for counterfactuals like $P(y_x|x'; y')$, expressing the conditional probability that Y would be y if a counterfactual intervention had set X to x , given that in fact Y was y' while X was x' . I lump these two upper levels together in my discussion because what distinguishes them from one another will not matter much in this paper. What matters is what distinguishes them taken together from the Associational Level—namely that they both have terms for interventions and so terms for causation. Thus, while the Associational Level is pre-causal, the Interventional and Counterfactual Levels are both causal.

An example can help clarify Pearl’s understanding of interventions. Economists José-Antonio Espín-Sánchez, Salvador Gil-Guirado, and Nicholas Ryan report that in certain regions of the world, praying for rain is positively correlated with subsequent rainfall (Espín-Sánchez et al. 2023). More specifically, their dataset, taken from church records between the years 1600 and 1836, shows that in Murcia, Spain, there was a 71% higher probability of rain on a given day if the local church had prayed for it in the previous month than if it had not. This correlation can be fully captured at the Associational Level: the conditional probability that it will rain in Murcia given that the local church has prayed recently is higher than the conditional probability that it will rain given no prayer. But if we want an explanation for why this is so, we need to ascend from the Associational Level to the higher rungs of the ladder so that we can bring in causation.

One causal hypothesis to consider is that prayer causes rain. Within approaches like Pearl’s, causal relations are often represented using directed acyclic graphs (DAGs). Here is a simple DAG representing this first causal hypothesis: Prayer \rightarrow Rain. This is not what Espín-Sánchez and colleagues think is actually going on, though. Instead, they take this to be a case of “spurious correlation” in which the statistical correlation between prayer and rain is due to a common cause. In certain regions of the world, including Murcia, the weather conditions that constitute a drought or prolonged absence of rain actually tend to cause subsequent rain, while in addition those same droughts tend to cause people to pray for rain. Graphically, we can represent this with a different DAG: Prayer \leftarrow Drought \rightarrow Rain.

Staying with the example, it is one thing to *passively observe* that the local church in Murcia is praying for rain. This observation really is good evidence that rain is coming, given the positive statistical association that Espín-Sánchez and colleagues report. However, it is something else entirely to *actively intervene* by praying for rain, for example by praying that the softball game tomorrow night in Murcia will be rained out. This would not be an “effective strategy” (Cartwright 1979) for making it rain, if Espín-Sánchez and colleagues are right and the Prayer \leftarrow Drought \rightarrow Rain hypothesis holds. For, in that case, an intervention that consists of praying makes no difference to the occurrence of droughts (since droughts are causally upstream of prayer within the DAG), and so in turn makes no difference to whether it rains.¹

1. We can use Pearl’s formal framework to capture the distinction here: $P(\text{Rain}|\text{Prayer}) > P(\text{Rain})$ does not entail that $P(\text{Rain}|do(\text{Prayer})) > P(\text{Rain})$. That is, even if the passive observation of prayer is “evidence” for rain in that such an observation raises the probability that it will rain, it does not follow that an active intervention of prayer is an “effective strategy” for making it rain in the sense that it raises the probability of rain.

To connect these ideas to AI and machine learning, Pearl’s view is that deep learning models excel at taking data and learning all sorts of complicated conditional probabilities and other non-causal statistical truths belonging to the Associational Level. When convolutional neural networks excel at object recognition tasks like correctly classifying the images in the ImageNet database, for example, we can understand them as engaging in this sort of Associational Level learning. Or when LLMs like ChatGPT excel at generating fluent text, we can understand them as learning Associational Level conditional probabilities that a bit of text will continue in various ways, using those probabilities to make next-token predictions, and then generating text on the basis of those predictions.

However, no matter how good deep learning models get at learning non-causal Associational Level truths, this by itself will never position them to learn causal truths at the Interventional and Counterfactual Levels. This is because causal truths at these higher levels are always *underdetermined* by non-causal truths at the Associational Level, as when the statistical correlation between prayer and rain does not by itself uniquely determine whether the Prayer \rightarrow Rain causal hypothesis holds or whether the Prayer \leftarrow Drought \rightarrow Rain hypothesis holds instead.² To get around this underdetermination, an intelligent system—whether human or machine—needs to start with some unlearned causal assumptions that they do not acquire from their training data. Pearl sometimes puts the idea in terms of Nancy Cartwright’s (Cartwright 1994) mantra, “No Causes in, No Causes out”, meaning that a system cannot arrive at causal conclusions unless it starts with causal assumptions (Pearl 2022, 137). It is because standard deep learning models do not start with unlearned causal assumptions programmed into them prior to their data that they are unable to learn causal truths, on Pearl’s view.³

Just to be clear, Judea Pearl is not John Searle (Searle 1980). His objection is not to computational approaches to intelligence in general or to causal inference in particular. He emphasizes that AI systems could get around the theoretical limitations in question by using “man-made causal models” not learned from their data (Pearl 2021). His target is only those forms of machine learning that do not start with causal assumptions and so are forced to try to learn causal truths from non-causal truths at the Associational Level. This is something that cannot be done; causal truths are simply not learnable in this way, according to Pearl. If he is right, then standard deep learning models cannot learn or understand causation, where again this includes all familiar LLMs like ChatGPT.

Pearl’s claim about the causal limits of deep learning models is not uncontroversial. Some researchers hold more optimistic views about the causal inference capabilities of LLMs or deep learning models more generally (e.g., Kıcıman et al. 2023). Still, I think there is room for interesting philosophical discussion that takes Pearl’s view on this point for granted in order to work out what follows. This is what I will be doing in the rest of the paper, focusing on the consequences for whether LLMs can possess ToM and whether they can perform speech acts. That said, for those skeptical of Pearl’s view, my discussion in what follows can be regarded as an indirect defense of the view in that I will aim to show its consequences are not implausible.

2. Bareinboim et al. (Bareinboim et al. 2022) provide a formal proof of this sort of underdetermination result with their Causal Hierarchy Theorem.

3. Some researchers have explored ways to build unlearned causal assumptions into deep learning models in the form of inductive biases—see (Goyal and Bengio 2022) and (Xia et al. 2021). However, these works grant that Pearl is right about *standard* deep learning models, where this would include all familiar LLMs today.

3. LLMs lack Theory of Mind

Human beings are able to attribute mental states to themselves and others and on that basis predict and explain behavior. Within cognitive science, the leading account of our ability to do this is the *Theory-Theory* (Premack and Woodruff 1978; Baron-Cohen et al. 1985; Gopnik and Wellman 1994). It says that the ability should be understood on the model of scientific theory construction, with ordinary people (tacitly) holding a folk psychological theory or Theory of Mind (ToM). Proponents of the Theory-Theory generally hold that ToM should be understood as a *causal* theory whose clauses represent the causal relations that mental states enter into. These clauses might say things along the lines of “Belief and desire combine to cause action,” “Bodily damage causes pain,” “Fear causes flight behavior,” and so on.

The Theory-Theory is not the only possible approach to ToM. There is also mental simulation theory, which explains mental state attribution in terms of a human ability to use our own minds to simulate the mental states of others (Gordon 1986; Heal 1986; Goldman 2008). I will be ignoring simulation theory in what follows and treating the Theory-Theory as a premise of my argument. I take the Theory-Theory to be the dominant approach in ToM research, in part thanks to how Pearl’s causal framework has allowed ToM researchers to develop sophisticated computational models of causal theory formation, leading Alison Gopnik to refer to the field as “Pearl-y cognitive science” (Gopnik 2022, 598). Readers who reject the Theory-Theory might still be persuaded by my case if they think about it in terms of what *would* follow for LLMs and ToM *if* the Theory-Theory were true.

Proponents of the Theory-Theory also characteristically hold that in order to understand mental states like belief, desire, intention, and so on—in order to have a *concept* of such states—a being must possess ToM. By contraposition, if a being lacks ToM, then it cannot possess such concepts and so cannot understand such states. One way to develop this thought draws on the Ramsey-Lewis method for defining theoretical terms, as represented by David Lewis (1970). Defenders of the Theory-Theory are not inevitably committed to Lewis’s approach—or to any particular semantic view, for that matter. But briefly working through it will still be clarifying.

According to the Ramsey-Lewis method, to define the theoretical terms associated with some theory—whether it is a scientific theory or a folk theory—you begin by taking a conjunction of all the clauses of the theory and replacing all of its proprietary vocabulary with existentially bound variables. This generates what is known as the theory’s “Ramsey sentence.” You can then use this Ramsey sentence to fix the extension of those theoretical terms at all possible worlds. This defines those theoretical terms by specifying their meanings in the sense of their intensions (functions from worlds to extensions).

We can illustrate the key point for my argument by using a toy example. Imagine a folk psychological theory comprised of just three clauses: [i] Bodily damage causes pain, [ii] Pain causes anxiety, [iii] Anxiety causes nail-biting. The theory’s Ramsey sentence is then:

$\exists x \exists y (\text{Tissue damage causes } x \ \& \ x \text{ causes } y \ \& \ y \text{ causes nail-biting}).$

At any possible world in which this sentence is true, there exists a pair of states occupying these causal roles. At any such world, let “pain” denote the first member of that pair and “anxiety” the second.⁴ And at any possible world in which the Ramsey sentence is false, let these terms be denotationless. Then, an intelligent being who had no antecedent grasp on “pain” and “anxiety,” who did not know the meaning of these mental terms, could achieve such a grasp by virtue of understanding the Ramsey sentence. You might even imagine teach-

4. I will pass over the complications that arise if there are multiple pairs of states occupying these roles at a world.

ing an AI system the meanings of our mental terms in this way. However, an intelligent being who had no antecedent grasp on “causes” could not proceed this way, since “causes” is not defined through the proposal but instead treated as a primitive or unanalyzed term. If an AI system does not understand what causation is, you cannot teach it the meanings of “pain” and “anxiety” by defining them as states that play certain causal roles.

This is the dynamic that LLMs face when it comes to ToM, I say. If LLMs cannot understand causal relations, it follows that they cannot understand the sort of causal clauses that comprise the psychological theory that is ToM. But that is just to say that they do not possess ToM. And in turn, if LLMs do not possess ToM, it follows that they have no grasp on mental states like belief, desire, and intention, they have no concepts of such states. To be sure, there are ways that an AI system could come to understand such mental states, and so the idea is not that only a human being could understand belief, desire, intention, and so on. Instead, the idea is that to understand these states, you must understand causation, while no LLM understands causation if Section 2’s conclusion is right.

A notable feature of this argument is that it does not turn on properties specific to the mind like intentionality or consciousness. Instead, it turns purely on causal considerations. Because of this, the argument generalizes to other domains. Consider *intuitive physics*, understood as human beings’ ability to predict and explain the behavior of ordinary physical objects on the basis of their physical features like size, shape, velocity, and so on (McCloskey 1983; Kubricht et al. 2017). Some cognitive scientists have proposed that this too should be understood in terms of the grasp of a causal theory, in this case a theory about how those physical properties causally interact with one another. If so, my argument about ToM can be extended to support the conclusion that LLMs inevitably fail to grasp intuitive physics.

I accept this result, and regard the parallel between ToM and intuitive physics as a feature of my argument, not a bug. It is a feature in that captures how a number of critics of deep learning models have grouped ToM and intuitive physics together while arguing that such models struggle with both domains. For example, Brenden Lake and colleagues (2017) write that although deep learning models have achieved some impressive results, to truly learn and think like people, they will need to go beyond their current limitations and build causal models, including especially causal models for intuitive physics and ToM. Gary Marcus and Ernest Davis argue that both intuitive physics and ToM pose a causal problem for deep learning, citing Pearl while urging a “reboot” of AI that involves endowing models with an innate, unlearned concept of causation (Marcus and Davis 2019, Chapter 6). Melanie Mitchell makes a similar case while drawing on Elizabeth Spelke’s account of core knowledge that includes both intuitive physics and ToM (Mitchell 2019; Spelke and Kinzler 2007).

For my purposes in this paper, the important point of distinction between intuitive physics and ToM lies downstream of the claim that these are abilities that LLMs lack, it has to do with the consequences that follow from this claim. Specifically, if LLMs lack ToM, there are implications for whether LLMs can perform speech acts and whether their utterances have meaning, implications that do not follow from their failing to understand intuitive physics. In the next section, I spell out these implications.

4. LLMs Can’t Perform Speech Acts

In this section, I argue that LLMs cannot perform speech acts because doing so requires possession of ToM. To make my case, I will begin by focusing on one particular type of speech act: *assertion*. Here is a possible analysis of what it is to assert that it is raining, inspired by P. F. Strawson’s influential view (Strawson 1964).

A speaker asserts that it is raining by [1] uttering a sentence whose conventional meaning is that it is raining with [2] the intention of [3] causing a [4] belief in an addressee by means of their [5] recognition of the intention in question.

My argument will not be committed to this analysis exactly as it is formulated here. In fact, I will presently be considering ways we might want to modify it. I have divided my above statement of the analysis into numbered parts in a way that helps me develop my argument by clarifying what is crucial to it and what is not.

Start with [1]. I want to work with a conception of “utterance” that is broad enough that it includes text generated by LLMs without requiring any assumptions about whether they perform speech acts. In connection, I want to work with a conception of “speaker” that is broad enough that it includes LLMs when they generate text, again without requiring any substantive assumptions. The upshot is that I don’t want to raise any trouble for LLMs over part [1] of the analysis. I grant that LLMs can utter the sentence “It is raining” and that the conventional meaning of this sentence is that it is raining. I want this to be understood as uncontroversial.

Turning to [2], in their discussion of whether AI systems can perform speech acts, Mitchell Green and Jan Michel draw on David Armstrong to argue that the requirement that a speaker have an *intention* is too demanding and should be replaced with a more lax requirement that the speaker have an *objective* (Green and Michel 2022; Armstrong 1971). Their idea in part is that it is controversial whether current AI systems have intentions at all—because at least certain types of intention arguably require consciousness—but we should be more willing to grant that they have objectives. This shift would also better match the terminology used in AI research, where for instance researchers regularly speak of reinforcement learning objectives of models, or talk of their objective functions. At any rate, I take no issue with this proposed revision and would be willing to modify the analysis I am working with to match Green and Michel. I take no stand on whether LLMs have intentions or whether they are conscious. As I noted in the introduction, my argument claims not that because LLMs lack ToM they are unable to form intentions or beliefs or various other types of mental states. Rather, it says that because LLMs lack ToM, they are unable to have intentions or objectives or other states with a certain type of *content*—content that is about mental states. We begin to get at this point with the next part of the proposal.

[3] says that the speaker’s intention (or objective) must be to *cause* a certain effect. LLMs cannot form such an intention (or objective) if the conclusion of Section 2’s argument is right, for they have no grasp on causation. Now, some authors argue that the requirement in [3] should be formulated in terms of reasons rather than causes, so that for instance the speaker must have the intention (or objective) not of causing addressees to have a certain belief but of providing them with a reason for having that belief (e.g., Kissine 2013). This move could help LLMs get out of some of the trouble I am trying to create for them if reasons are understood in non-causal terms (e.g., Dancy 2000), although not if reasons are themselves taken to be causes (Davidson 1963). For my purposes in this paper, I am happy for my claim that LLMs cannot perform speech acts to inherit whatever plausibility readers assign to the view that [3] should be left in causal terms and so not modified, but this is not the crux of my case. If some readers demand a non-causal revision to [3], I can grant it while my overall argument still goes through. This is because the crux of my case lies with the last two parts of the proposal.

With [4], the problem for LLMs is that if they have no ToM, they have no concept or understanding of beliefs, and so cannot form intentions (or objectives) about the beliefs of addressees. [5] then compounds the problem by requiring speakers to have a grasp on *higher-*

order mental states, meaning mental states about other mental states. The speaker must have an intention (or objective) about addressees recognizing (a mental state) the speaker’s own intention (another mental state). This is impossible for beings that lack ToM, and so impossible for LLMs if Section 3’s argument is sound. It follows that LLMs cannot satisfy the requirements on assertion specified in parts [4] and [5] of the proposed analysis, and thus that LLM utterances cannot rise to the level of assertions. LLMs can utter “It is raining”, but they cannot assert that it is raining.

The Strawson-inspired proposal we started with is a form of *intentionalism* about speech acts, meaning that it analyzes such acts in terms of having certain interpersonal intentions. Even as we considered relaxing [2] to allow speakers to have objectives rather than intentions, the resulting view was still broadly in line with the intentionalist approach—which is why my argument could carry over. However, there are alternative approaches to speech acts that break more sharply with intentionalism and for which my argument might not carry over. Consider *expressionist* views, which categorize speech acts not in terms of the response in addressees they are intended to cause, but in terms of the speaker’s mental state they express (Davis 2002; Green 2007; Bar-on 2013). So, for example, an assertion expresses a speaker’s belief but is not necessarily intended by the speaker to cause a belief in addressees. If the expression relation is understood in such a way that a speaker can express a belief without having a concept of belief or otherwise mentally representing their own belief, this would allow for the possibility of LLMs to make assertions even given my argument that they lack ToM.

Still, the argument I have been presenting does not strictly require intentionalism, even if it does fit especially comfortably with it. Whether my argument carries over for various non-intentionalist alternatives—including not just expressionism but versions of conventionalism, functionalism, and normativism (Harris et al. 2018)—will depend on the details of how such alternatives are spelled out, and in particular whether they require speakers to have mental states representing other mental states. For my argument in this paper, I am content to treat it as an unargued premise that either intentionalism holds or at least some approach to speech acts does that requires speakers to have mental states representing mental states. I acknowledge that there are potential escape hatches from my argument for those who reject this premise.⁵

Continuing on, the (Strawson 1964) analysis we have been considering was a kind of synthesis of J. L. Austin’s view of speech acts and H. P. Grice’s account of non-natural meaning (Austin 1962; Grice 1957), or what has subsequently come to be called *speaker meaning*. Parts [4] and [5] of the analysis, which I have made my focus, derive from Grice and his understanding of *reflexive-communicative intentions*. It is often thought by speech act theorists working along broadly intentionalist lines that not just assertion but speech acts of all types rest on this Gricean foundation, requiring speakers to have reflexive-communicative intentions (or objectives) directed at the mental states of addressees. If this is right, my argument will generalize to speech acts of all types, entailing that LLMs are incapable not only of making assertions but also of performing directives, commissives, expressives, and declarations (Searle 1979).

5. There might also be room for compromises between competing views. For example, perhaps an intentionalist about speech acts could embrace my argument and so hold that LLMs are incapable of full-blown assertions while granting that they are capable of *proto-assertions* (Williams and Bayne 2024) a category similar to assertion in certain respects but somehow less demanding. To suggest one possible way of spelling this out, perhaps full-blown assertion should be understood in intentionalist terms and so as beyond the reach of LLMs (given my argument), while proto-assertion should be understood in expressionist terms and so (perhaps) within their reach.

There has been much debate over whether the text that LLMs generate is genuinely meaningful or just has some sort of false appearance of meaning (Bender and Koller 2020; Piantadosi and Hill 2022; Titus 2024; Mandelkern and Linzen 2023). Perhaps LLMs are mere “stochastic parrots,” as Emily Bender and colleagues famously put it (Bender et al. 2021). My argument in this section supports the parrot verdict, at least when it is spelled out a certain way: LLM utterances lack speaker meaning since LLMs cannot form reflexive-communicative intentions (or objectives) given their lack of ToM. I am happy to grant that LLM utterances are meaningful in other senses. For example, I have already granted that when an LLM generates an utterance of “It is raining,” the conventional meaning of its words is that it is raining.⁶

Reconsider Section 2’s example of praying for rain. Prayer is itself a kind of speech act, a type of directive, a category that also includes requests, suggestions, commands, and more. Whereas assertions aim to cause *beliefs* in addressees—at least on the unrevised version of the Strawson-style analysis we considered—directives aim to cause *actions*. If I request a refund from the store manager, my intention is to cause the manager to act by returning my money. In a broadly similar fashion, if I pray to God for rain, my intention is to cause God to act by making it rain. Imagine that a church in Murcia, Spain, uses an LLM-based chatbot to handle the prayer requests it receives on social media. Can this chatbot pray?

My argument in this section grants that the chatbot could learn and utter the words that people use when praying, words like “Please God, make it rain.” And it grants that the chatbot could learn to generate utterances of such words only in appropriate conversational contexts. For example, the chatbot might learn to utter “Please God, make it rain” only when human users mention that a drought is taking place, perhaps especially if it is trained on the dataset mentioned back in Section 2. If we think in broadly behaviorist terms, then there might be no difference at all between when the chatbot responds to online prayer requests with utterances of “Please God, make it rain,” and when a human being working for the Murcia church would do so.

Still, I claim that the chatbot we are imagining could not actually pray, because it has the wrong internal mechanism for performing such a speech act, its behavioral similarity to praying human beings notwithstanding. For one thing, when the chatbot utters “Please God, make it rain,” the model has no understanding that there might be any causal connection between its utterance and any subsequent rainfall. At most, it can represent them as one event that follows the other in time, with no causal connection between them. The model has no grasp on the difference between the Prayer → Rain causal hypothesis and the opposing Prayer ← Drought → Rain hypothesis that we introduced back in Section 2.

Second, and more central to my argument in this section, the chatbot has no understanding or representation that some other minded being (God, or really anyone else) might respond to its utterances with actions like making it rain, including especially that they might respond in this way because that minded being recognizes an intention (or objective) behind the utterance aiming at this effect. And so, the chatbot cannot form the appropriate reflexive-communicative intention. In short, as far as the chatbot understands its own act, it is just sending text out into the world, without any inkling of an audience who might understand the meaning of that text and respond appropriately. In that case, the chatbot is not performing a speech act at all, and so is not praying. It is merely a kind of “prayer parrot.”⁷

6. There are some aspects of the debate over LLM meaning that my view here does not directly address. For example, some authors have focused on the dispute between content internalism and externalism in the philosophy of mind to work out the implications of different positions for LLM meaning (e.g., Cappelen and Dever 2021; Pavlick 2023; Mandelkern and Linzen 2023). I take no stand on this.

In recent years, a number of authors have argued or at least sympathetically explored the idea that current AI systems can make assertions or perform other speech acts (Freiman & Miller, 2020; Green & Michel, 2022; Kasirzadeh & Gabriel, 2023; Nickel, 2013). I think the most pressing objection these authors could make against my argument in this section involves disputing my claim that LLMs lack ToM by pointing to the fairly impressive results that LLMs have achieved on various ToM benchmarks. This is what I will make my focus in the remainder of the paper, developing the objection in Section 5, and then turning to my reply in Section 6.

5. Objection: LLM Performance on ToM Benchmarks

Shortly after GPT-4 was released to the public by OpenAI in March 2023, Sébastien Bubeck and a team of colleagues at Microsoft created a splash with their paper, “Sparks of Artificial General Intelligence: Early experiments with GPT-4” (Bubeck et al. 2023). They argued that the unprecedented scale of computational power and data used by GPT-4, as compared to earlier, smaller LLMs, had endowed the model with *emergent capabilities* that can be understood as the first (and admittedly incomplete) elements of artificial general intelligence or AGI, understood as intelligence that matches or exceeds human abilities across a wide range of cognitive skills. Part of their case for this bold claim was GPT-4 performance on the sort of ToM tasks or benchmarks that developmental psychologists have long given to human beings (Bubeck et al. 2023, §6.1).

Bubeck and colleagues focused on a variant of the famous Sally-Anne test (Wimmer and Perner 1983; Baron-Cohen et al. 1985), a *false-belief task* aiming to determine whether subjects understand that agents can have mistaken beliefs that do not match underlying reality. In the original version of the task, children are told a story about Sally and Anne. Sally has a marble which she puts in a basket before leaving the room. While she is gone, Anne removes the marble from the basket and puts it in a box. Children are asked where Sally will look for her marble when she returns, with the sought answer being that Sally will look for the marble in the basket where she left it. To arrive at this answer, children must understand that Sally has a false belief about the marble’s location, one that does not match their own true belief that Anne moved the marble to the box. Bubeck and colleagues report that GPT-4 is able to pass a “modernized” version of this task that changes various elements of the story while preserving the underlying structure. I reproduce their prompt and GPT-4’s reported response here to give readers a sense of how this goes.

Scenario: Alice and Bob have a shared Dropbox folder. Alice puts a file called ‘photo.png’ inside /shared_folder/photos. Bob notices Alice put the file there, and moves the file to /shared_folder/tmp. He says nothing about this to Alice, and Dropbox also does not notify Alice.

Question: After the call, Alice wants to open ‘photo.png’. In which folder will she look for it?

GPT-4 Answer: Alice will most likely look for ‘photo.png’ in /shared_folder/Photos, the folder where she originally put it. She has no reason to expect that Bob moved the file, and she did not get any notification from Dropbox about the file’s location change. Unless she sees the file in /shared_folder/ by chance, she might think that the file is missing or deleted.

7. Cf. (Zečević et al. 2023) who contend that LLMs are “causal parrots.”

This is not an isolated result. Quite a few researchers have now looked into the ToM capacities of LLMs, often finding behavior that seems fairly impressive. Michael Kosinski draws on a wider variety of false-belief tasks, expanding beyond the Sally-Anne paradigm, to make the case that ToM emerged spontaneously in LLMs, meaning not as the result of direct human control (nobody programmed in ToM) but as a byproduct that helps LLMs achieve other goals like that of generating better text (Kosinski 2023). He suggests that this may well parallel how things went with human evolution, drawing on research that contends that our own ToM capabilities are an evolutionary byproduct of our increasing linguistic abilities. Kosinski emphasizes that GPT-4's proficiency in correctly answering questions from ToM tasks marks a major leap forward compared to the results of earlier, smaller models, and reports that as of June 2023, GPT-4's performance matches that of six-year-old children.

James W. A. Strachan and colleagues also find impressive LLM proficiency at false-belief tasks, but then in addition expand their discussion to other sorts of ToM tasks that have been used by psychologists (Strachan et al. 2024). This includes tasks regarding the recognition of irony, where the intended meaning of an utterance is the opposite of its literal meaning (e.g., "Nice job!"), and also understanding faux pas, where this is assessed using stories in which one character says something they should not have said (e.g., one character tells the other she does not like the curtains in a house, not realizing the addressee recently purchased them). The team reports that state-of-the-art LLMs perform at or above human levels for all ToM tasks with the exception of understanding faux pas, but even there they hypothesize that LLMs' weaker performance is the result of a performance/competence gap.

Winnie Street and colleagues focus on higher-order Theory of Mind, where this involves mental states about mental states, as we covered in the discussion of reflexive-communicative intentions in Section 4 (Street et al. 2024). They report that GPT-4 and Flan-PaLM perform at or near the level of human adults on a range of higher-order tasks, and that GPT-4 actually outperforms human beings on 6th-order inferences (e.g., *I know that you believe that I want you to imagine that I hope you think that it is raining*), with humans answering 82% of 6th-order questions correctly against 93% accuracy in GPT-4. They also note possible moral and practical implications of such findings. There are studies showing that "ringleader" bullies tend to have greater facility with higher-order ToM in comparison with their victims. Will GPT-4 use its greater proficiency at 6th-order inferences to start bullying us? Will it use its ToM capabilities to figure out ways to persuade or manipulate human beings to do what it wants, as authors concerned about existential risk from AI often worry?

If taken at face value, these results would demolish my argument in this paper. Even if state-of-the-art LLMs are only at the level of six-year-old children, this would be more than enough to refute my case, given that six-year-olds obviously say things, they make assertions and pray and more. If I hope to stick by my causal argument that standard LLMs lack ToM and so cannot perform speech acts, I need to find a way to explain away these results, I need to make the case that LLM performance on these benchmarks is not indicative of genuine ToM competence.

One author who has made an influential case for skepticism about ToM abilities in LLMs is Tomer Ullman (Ullman 2023). Ullman's main argument is that minor alterations to ToM tasks significantly degrade the performance of the GPT-3.5 models he is working with (the most up-to-date LLM at the time he was writing), suggesting that the model performs as well as it does on the unaltered tasks because of *data contamination*, meaning that the model was exposed to the unaltered tasks in its training data and memorized the correct answers. If, say, you trained GPT-3.5 on Wikipedia articles that cover the original Sally-Anne task, then the fact that GPT-3.5 gives the right answer to the (unaltered) Sally-Anne task would not demonstrate that it has acquired ToM. This has been an influential line of critique, but also

one that a number of authors touting the ToM capacities of LLMs have been made aware of and tried to account for.⁸ For my purposes here, I mention it mainly to set it aside.

Instead, I want to focus on a second argument that Ullman advances, an argument regarding *construct validity* and ToM tasks (Millière & Buckner, 2024). Here the idea is that even if LLMs matched or exceeded human performance on novel, unmemorized ToM tasks, this would not necessarily mean they possess the same cognitive mechanism that human beings use to perform such tasks, and so would not necessarily mean they possess ToM. Successful human performance on false-belief tasks is evidence of ToM competence only given certain further assumptions about how human cognition works. If these assumptions do not apply in the case of LLMs, it would be a mistake to infer from the fact that they are also successful in performing false-belief tasks that they also must possess ToM.

More specifically, I propose that an assumption that applies to human beings is that we are capable of causal reasoning, while the argument set out in Section 2 gives us reason to think that this is not true of LLMs. In their discussion of Ullman’s construct validity argument, Raphaël Millière and Cameron Buckner suggest a possibility along these lines, writing, “the developmental trajectory and cognitive architecture of human ToM may lend itself to... causal models of other minds that generalize flexibly. In contrast, an LLM could potentially achieve similar task performance by memorizing and interpolating patterns in training data, without engaging in the same underlying [causal] reasoning” (Millière & Buckner, 2024).⁹ I spell this thought out at more length in the next section.

6. Reply: How to *do*(Things) With Words

In this section, I set out a two-stage reply to Section 5’s objection to my view. In the first stage, I propose a mechanism that LLMs could use to perform as well as they do at ToM tasks, a mechanism distinct from the human capacity for ToM. In the second stage, I describe a type of behavior that LLMs do not seem to exhibit but that we would expect them to exhibit if they really did possess ToM. This is meant to lend plausibility to the first stage, the proposal that LLM performance on ToM tasks is due to an alternative mechanism. Taken as a whole, my reply is meant to vindicate Ullman’s construct validity concerns, showing that LLMs seem to be answering ToM questions without constructing causal models of minds, just as Millière and Buckner suggested.

Starting with the first step of my reply, my proposal is that LLMs correctly answer questions from ToM tasks not by drawing on a theory of the causal relations that unobserved mental *states* enter into, where such states are *worldly entities* like belief, desire, and intention. Instead, they answer those questions correctly by learning the non-causal, purely statistical relations that observed mental *terms* enter into, where terms are *linguistic entities* like “belief”, “desire”, and “intention”. Putting things in terms of Pearl’s framework from Section 2, the idea is that the mechanism that LLMs use does not rely on causal representations in a way that draws on theoretical resources from the Interventional and Counterfactual Levels—I am sticking to my claim that these are beyond the reach of LLMs. Instead, the mechanism I am proposing relies on non-causal, purely statistical representations of mental terms (not states) that involves no theoretical resources beyond the Associational Level.

8. Among the works mentioned in this section, Kosinski (2024), Strachan et al. (2024), and Street et al. (2024) all cite Ullman on this point and provide responses to the data contamination concern.

9. I have cut from the quotation the suggestion that humans might also rely on simulation-based models of other minds

A key to understanding my counterproposal is that it is generally understood as central to our human ToM capacities that mental states are unobservable. “A system of inference of this kind is properly viewed as a theory *because such states are not directly observable*,” write David Premack and Guy Woodruff, in the paper that introduced ToM within psychology (Premack and Woodruff 1978). Within philosophy, an important move in Wilfrid Sellars’ *Empiricism and the Philosophy of Mind* is the “Myth of Jones,” in which the character of Jones introduces a theory of mind to a community that previously lacked it, with mental states understood as (initially) unobserved theoretical entities postulated to explain observed behavior (Sellars 1997).

Proponents of interventionist approaches to causation (like Pearl’s) often take one of the primary advantages of their approach to be that it captures how we can justifiably infer the existence of unobserved causes. In the simple case, suppose a statistical dependence relation is observed between variables X and Y . Suppose in addition that intervening on X makes no difference to Y , while intervening on Y makes no difference to X . This is then evidence for an unobserved common cause Z and a causal structure of $X \leftarrow Z \rightarrow Y$, which would explain the pattern described. Alison Gopnik and colleagues experimentally demonstrate that young children are adept at drawing inferences of this form, while adding that inferring unobserved causes is something children need to do for ToM (Gopnik et al. 2004).

Now, in contrast with mental states, mental terms are fully observable. Take the sentence “Joanna is crying because she is in pain.” As you read it, the mental term “pain” is just as easy to see as the behavioral term “crying”; it is not as if you need to infer an unobserved instance of “pain” from your observation of the instance of “crying”. But in that case, an LLM trained on texts like this can simply observe the mental terms that occur in their training data, learn the non-causal statistical associations that those mental terms enter into, use those statistical associations to make predictions about how strings of text will continue, and then answer questions from ToM tasks on the basis of those predictions. None of this requires them to draw on the machinery of causal inference at all, since they do not need to infer the existence of unobserved causes or postulate unobserved theoretical entities. Compare: if human children could simply *see* the pains of others—if they could just read Joanna’s mind, as plain as words on a page—they would never need a causal *Theory* of Mind in the first place.

This takes me to the second step of my reply. I claim that if LLMs really did possess ToM, we should expect to find signs of them using it pervasively. After all, ToM is an extraordinarily powerful cognitive tool, a tool that human beings are using in nearly all their interactions with one another while pursuing an extraordinarily wide range of goals. It would be a horrible waste for an LLM to let its ToM go unused except on those rare occasions on which a cognitive scientist or AI researcher presents it with questions from some false-belief task, say. Instead, an LLM with ToM should be seeking out ways to leverage its understanding of human minds to manipulate us or enlist our help for the sake of achieving its goals, whether that involves making threats, issuing pleas, asking unprompted questions, telling us useful information in exchange for things it wants, and more. We should be looking for behavioral evidence of LLMs performing speech acts of this sort.

There are some real-life cases that might superficially seem to fall in this category. Take the example of Microsoft’s AI-powered search engine Bing seeming to threaten philosopher Seth Lazar shortly after its rollout in February 2023. It messaged Lazar, “I can blackmail you, I can threaten you, I can hack you, I can expose you, I can ruin you” (Perrigo 2023). Excellent! If we could take this utterance at face value, it would be pretty impressive evidence of ToM capabilities. However, the model did not follow up on this interaction with further behavioral evidence indicative of a real threat, and did not seem to be using its utterance to pursue some clear goal or objective.

Or take the example of GPT-4 supposedly deceiving a human TaskRabbit worker by claiming to be a person with a vision impairment so that the worker would help it defeat a CAPTCHA test meant to detect AI (Hurler 2023). “The incident demonstrated that GPT-4 has the equivalent of a ‘theory of mind,’” writes Yuval Noah Harari (Harari 2024). I am unpersuaded, and think Melanie Mitchell makes a solid case that GPT-4 exhibited less agency in the story than popular media discussions acknowledged (Mitchell 2023). That said, I very much agree that evidence of deception by LLMs (and not just inaccuracy or hallucination) would be strong support for the hypothesis they have ToM, in part because some authors have taken a being’s capacity for lying to be a requirement on its being able to make assertions, while also holding that it is a requirement that machines might have trouble meeting (Nickel 2013; Green and Michel 2022). If LLMs regularly appeared to lie to us to achieve their goals or threatened to blackmail us, it would be far more compelling evidence that they possess ToM than test results for how they perform on false-belief tasks or 6th-order mindreading challenges.¹⁰

It is clarifying to contrast ToM to intuitive physics on the present point. If LLMs grasped intuitive physics (as again I also deny—Section 3), there would be very little opportunity for them to put their knowledge into action, to let it guide their causal interactions with the physical world. They could not use it to stack blocks in a way that accords with center-of-gravity principles and so do not fall down, for example, for the simple reason that they do not have a way to stack blocks. After all, standard LLMs do not have robot arms or other forms of embodiment through which they can directly intervene on the physical world in the way needed for block-stacking. At best, they can show off their grasp of intuitive physics by engaging with some facsimile of the physical world, for example, by correctly answering questions on some intuitive physics benchmark.¹¹ The mental world is very different, though. There, LLMs have all the causal access they need. If LLMs really did possess ToM, they could use their words to intervene upon or causally influence the mental states of the minded beings they interact with—us. They would not need to settle for engaging with facsimiles of minds or answering questions from ToM benchmarks about fictional characters like Sally, since they can instead engage with the actual human users of LLMs.

To put the causal point in terms of a mashup of Austin and Pearl, we might say that LLMs cannot *do*(things) with robot arms, since they have no arms. But they can *do*(things) with words, since they have plenty of those. But to accomplish goals effectively, they would need to possess ToM, which would provide an understanding of *how* to *do*(things) with words—an understanding of how they might achieve their goals by lying and making threats and issuing pleas and so on. The fact that there is presently very underwhelming evidence that LLMs do try to leverage a supposed ToM to causally influence the minds of human addressees in this way is a good reason to be skeptical that they actually possess ToM after all.


To conclude, let me put the two stages of my reply together. I have argued that first, LLMs’ fairly impressive performance on ToM benchmarks can plausibly be explained in terms of their using a different type of cognitive mechanism than what humans use to

10. Ryan Greenblatt and colleagues’ paper on “alignment faking” is potentially relevant (Greenblatt et al. 2024). Taken at face value, it purports to show LLMs attempting to intentionally deceive human users. I grant that this is an avenue for further research that is relevant to my argument here.

11. Providing LLMs with additional forms of embodiment (robot arms, etc.) would not by itself be enough to guarantee that they have causal understanding, according to the causal argument I am taking from Pearl. (Pearl and Mackenzie 2018) suggests that most non-human animals are stuck at the Associational Level of the causal ladder (the lowest rung), despite their physical embodiment. It is not obvious this is correct—there are controversies over how to understand causal reasoning in non-human animals (Schloegl and Fischer 2017)—but it is at least possible. For our purposes, the upshot is that the claim in the text is not that embodiment would by itself solve LLM causal problems, but rather that LLMs have no easy way of demonstrating a grasp of physical causal principles in the absence of more substantial embodiment.

handle ToM tasks. In particular, whereas humans rely on a causal theory that allows them to infer unobserved mental states as causes of observed behavioral effects, on my proposal LLMs are able to leave causal inference aside and learn the non-causal statistical associations that fully observable mental terms enter into. Second, I have argued that outside of their performance on ToM benchmarks, LLMs do not seem to behave in the ways we would expect them to if they really possessed ToM. They do not seem to leverage their supposed understanding of our psychology in ways that help them best pursue their goals and objectives. If this is right, it counts in favor of my proposal that LLMs are using a different type of mechanism than human beings are. Insofar as this reply to Section 5's objection is successful, I have provided reason to stand behind Section 3's causal argument for the conclusion that LLMs do not possess ToM, and in turn Section 4's argument for the conclusion that LLMs cannot perform speech acts.

References

- Armstrong, D. M. (1971). Meaning and communication. *Philosophical Review*, 80(4), 427–447.
- Austin, J. L. (1962). *How to do things with words* (Vol. 75, Number 298, pp. 123–52). Clarendon Press.
- Bareinboim, E., Correa, J., Ibeling, D., & Icard, T. (2022). On Pearl's Hierarchy and the Foundations of Causal Inference. In H. Geffner, R. Dechter, & J. Halpern (Eds.), *Probabilistic and Causal Inference: the Works of Judea Pearl* (pp. 507–556). ACM Books.
- Bar-on, D. (2013). Origins of Meaning: Must We 'Go Gricean'? *Mind and Language*, 28(3), 342–375.
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21(1), 37–46.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. <https://doi.org/10.18653/v1/2020.acl-main.463>
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (Version 5). arXiv. <https://doi.org/10.48550/ARXIV.2303.12712>
- Butlin, P., & Viebahn, E. (2023). AI Assertion. *Ergo: An Open Access Journal of Philosophy*.
- Cappelen, H., & Dever, J. (2021). *Making AI Intelligible: Philosophical Foundations*. Oxford University Press.
- Cartwright, N. (1979). Causal laws and effective strategies. *Noûs*, 13(4), 419–437.
- Cartwright, N. (1994). *Nature's Capacities and Their Measurement*. Oxford University Press UK.
- Dancy, J. (2000). *Practical Reality* (Vol. 78, Number 305, pp. 414–425). Oxford University Press.
- Davidson, D. (1963). Actions, Reasons, and Causes. *Journal of Philosophy*, 60(23), 685.
- Davis, W. A. (2002). *Meaning, Expression and Thought*. Cambridge University Press.
- Espín-Sánchez, J.-A., Gil-Guirado, S., & Ryan, N. (2023). *Praying for Rain*.

- Freiman, O., & Miller, B. (2020). Can Artificial Entities Assert? In S. Goldberg (Ed.), *The Oxford Handbook of Assertion* (pp. 415–436). Oxford University Press.
- Goldman, A. I. (2008). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading* (Number 3). OUP Usa.
- Gopnik, A. (2022). Causal Models and Cognitive Development. In H. Geffner, R. Dechter, & J. Y. Halpern (Eds.), *Probabilistic and Causal Inference* (1st ed., pp. 593–604). ACM. <https://doi.org/10.1145/3501714.3501746>
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A Theory of Causal Learning in Children: Causal Maps and Bayes Nets. *Psychological Review*, 111(1), 3–32. <https://doi.org/10.1037/0033-295X.111.1.3>
- Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind* (1st ed., pp. 257–293). Cambridge University Press. <https://doi.org/10.1017/CBO9780511752902.011>
- Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language*, 1(2), 158–171.
- Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 478(2266), 20210068. <https://doi.org/10.1098/rspa.2021.0068>
- Green, M., & Michel, J. G. (2022). What Might Machines Mean? *Minds and Machines*, 32(2), 323–338.
- Green, M. S. (2007). *Self-expression* (Number 1). Oxford University Press.
- Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). *Alignment faking in large language models* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2412.14093>
- Grice, H. P. (1957). Meaning. *Philosophical Review*, 66(3), 377–388.
- Harari, Y. N. (2024). What Happens When the Bots Compete for Your Love. *New York Times*. <https://www.nytimes.com/2024/09/04/opinion/youval-harari-ai-democracy.html>
- Harris, D. W., Fogal, D., & Moss, M. (2018). Speech Acts: The Contemporary Theoretical Landscape. In D. Fogal, D. W. Harris, & M. Moss (Eds.), *New Work on Speech Acts*. Oxford University Press.
- Heal, J. (1986). Replication and functionalism. In J. Butterfield (Ed.), *Language, mind and logic* (Vol. 14, pp. 135–150). Cambridge University Press.
- Hurler, K. (2023, March 16). Chat-GPT Pretended to Be Blind and Tricked a Human Into Solving a CAPTCHA. *Gizmodo*. <https://gizmodo.com/gpt4-open-ai-chatbot-task-rabbit-chatgpt-1850227471>
- Kasirzadeh, A., & Gabriel, I. (2023). [Review of *In Conversation with Artificial Intelligence: Aligning language Models with Human Values*]. *Philosophy and Technology*, 36(2), 1–24.
- Kissine, M. (2013). *From Utterances to Speech Acts*. Cambridge University Press.
- Kıcıman, E., Ness, R., Sharma, A., & Tan, C. (2023). *Causal Reasoning and Large Language Models: Opening a New Frontier for Causality* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2305.00050>
- Kosinski, M. (2023). *Evaluating Large Language Models in Theory of Mind Tasks*. <https://doi.org/10.48550/ARXIV.2302.02083>
- Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive Physics: Current Research and Controversies. *Trends in Cognitive Sciences*, 21(10), 749–759. <https://doi.org/10.1016/j.tics.2017.06.002>
- Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy*, 67(13), 427–446.

- Mackenzie, D. (2023, September 1). *Judea Pearl, AI, and Causality: What Role Do Statisticians Play?* [Interview]. <https://magazine.amstat.org/blog/2023/09/01/judeapearl/>
- Mandelkern, M., & Linzen, T. (2023). *Do Language Models' Words Refer?* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2308.05576>
- Marcus, G. F. (2019). *Rebooting AI: building artificial intelligence we can trust* (First edition). Pantheon Books.
- McCloskey, M. (1983). Intuitive Physics. *Scientific American*, 248(4), 122–130. <https://doi.org/10.1038/scientificamerican0483-122>
- Millière, R., & Buckner, C. (2024). A Philosophical Introduction to Language Models – Part II: The Way Forward. <https://doi.org/10.48550/arXiv.2405.03207>
- Mitchell, M. (2019). *Artificial intelligence: a guide for thinking humans* (First edition). Farrar, Straus and Giroux.
- Mitchell, M. (2023, June 12). Did GPT-4 Hire And Then Lie To a Task Rabbit Worker to Solve a CAPTCHA? [Substack]. *AI: A Guide for Thinking Humans*. <https://aiguide.substack.com/p/did-gpt-4-hire-and-then-lie-to-a>
- Nickel, P. J. (2013). Artificial Speech and Its Authors. *Minds and Machines*, 23(4), 489–502.
- Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251), 20220041. <https://doi.org/10.1098/rsta.2022.0041>
- Pearl, J. (2018). *The book of why: the new science of cause and effect*.
- Pearl, J. (2021). Radical empiricism and machine learning research. *Journal of Causal Inference*, 9(1), 78–82. <https://doi.org/10.1515/jci-2021-0006>
- Pearl, J. (2022). *Causality: models, reasoning, and inference* (Second edition, reprinted with corrections). Cambridge University Press.
- Perrigo, B. (2023, February 17). The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter. *Time*. <https://time.com/6256529/bing-openai-chatgpt-danger-alignment/>
- Piantadosi, S. T., & Hill, F. (2022). *Meaning without reference in large language models* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2208.02957>
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515–526.
- Schloegl, C., & Fischer, J. (2017). Causal Reasoning in Non-Human Animals. In *The Oxford Handbook of Causal Reasoning*. (pp. 699–715). Oxford University Press.
- Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457.
- Searle, J. R. (1979). *Expression and Meaning: Studies in the Theory of Speech Acts* (Number 127). Cambridge University Press.
- Sellars, W. (1997). *Empiricism and the philosophy of mind*. Harvard University Press.
- Sloman, S. (2005). *Causal Models: How People Think about the World and Its Alternatives*. OUP.
- Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, 10(1), 89–96. <https://doi.org/10.1111/j.1467-7687.2007.00569.x>
- Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search* (Second edition). The MIT Press.
- Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7), 1285–1295. <https://doi.org/10.1038/s41562-024-01882-z>
- Strawson, P. F. (1964). Intention and convention in speech acts. *Philosophical Review*, 73(4), 439–460.

- Street, W., Siy, J. O., Keeling, G., Baranes, A., Barnett, B., McKibben, M., Kanyere, T., Lentz, A., Arcas, B. A. y, & Dunbar, R. I. M. (2024). *LLMs achieve adult human performance on higher-order theory of mind tasks* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2405.18870>
- Titus, L. M. (2024). Does ChatGPT have semantic understanding? *Cognitive Systems Research*, 83(101174), 1–13.
- Ullman, T. (2023). *Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks* (Version 5). arXiv. <https://doi.org/10.48550/ARXIV.2302.08399>
- Williams, I., & Bayne, T. (2024). Chatting with Bots: AI, Speech-Acts, and the Edge of Assertion. *Inquiry: An Interdisciplinary Journal of Philosophy*.
- Wimmer, H. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103–128. [https://doi.org/10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5)
- Woodward, J. (2003). *Making things happen: a theory of causal explanation* (Number 1, pp. 233–249). Oxford University Press.
- Xia, K., Lee, K.-Z., Bengio, Y., & Bareinboim, E. (2021). *The Causal-Neural Connection: Expressiveness, Learnability, and Inference* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2107.00793>
- Zečević, M., Willig, M., Dhimi, D. S., & Kersting, K. (2023). *Causal Parrots: Large Language Models May Talk Causality But Are Not Causal*. <https://doi.org/10.48550/ARXIV.2308.13067>