Conversations with Chatbots[1]

Patrick Connolly

This is a late draft, please cite published version to appear in Conversations Online: Explorations in Philosophy of Language, (eds. Connolly, Goldberg, Saul), OUP

The problem I consider in this chapter emerges from the tension we find when we look at the design and architecture of chatbots on the one hand and consider their conversational aptitude on the other. In the way that LLM chatbots are designed and built, there seems no good reason to suppose they possess second-order capacities such as intention, belief or knowledge. Yet we have developed theories of conversation that make great use of second-order capacities of speakers and their audiences to explain how aspects of conversation succeed. As we can all bear witness to now though, at the point of use chatbots appear capable of performing language tasks at a level close to that of humans. This creates a tension when we consider something like, for example, the classic Gricean theory of implicature. On a broad summary of this type of account, to utter *p* and implicate *q* requires the reflexive occurrence of an audience supposing a speaker believes that *q*, and the speaker believing that their audience can determine they believe it when they utter *p*. So taken at face value, if a chatbot doesn't have the capacity for belief, then either in their role as speaker or audience, they would not seem capable of either generating or comprehending implicatures. As will be shown later, though, on the surface it does seem that chatbots are capable of dealing with (some) implicatures, and as such it raises questions about how we should then correlate this with what we think occurs in cases of implicature with chatbots.

[1] Many thanks are owed to Jenny Saul for early discussions on this and to Sandy Goldberg and Jenny for very helpful feedback on an earlier draft. Also thanks to the members of the Digital Pragmatics and Epistemology Reading Group where I've been shaping these ideas.

Here's a summary of how the chapter goes. In Section 1 I explain what I mean by 'chatbot', specifically I give some brief details on how the current state-of-the-art Large Language Model (LLM) powered chatbots function. The aim here is to motivate the conjecture that despite being impressive feats of engineering, these types of machines, at the level of design and build, don't appear to have intentions, beliefs or knowledge.[2] In section 2 I consider evidence that suggests chatbots can perform many linguistic tasks to a near-human-like ability. In considering my own interactions with chatbots and by looking at some of the nascent empirical work on their linguistic performance, I argue that there is a good case to be made that machines are sufficiently proficient at the use of natural language that we might expect our language theories should be able to accommodate machine usage. In section 3 I explain how these conjectures create problems for conversational theory with a particular focus on conversational implicature and conclude by briefly looking at potential ways we might seek to avoid the problem.

1 Chatbots: Design and architecture

In this section, I set out some terminology, provide a brief overview of what chatbots are, and a rough explanation of how state-of-the-art versions work at the time of writing. Overall, I want to give weight to the conjecture that in terms of how they're built and the types of tasks they perform, chatbots aren't particularly mysterious entities. They are sophisticated works of technical engineering that are in many ways astonishing, but ultimately their design and architecture, and the tasks they do to process and generate language, gives us no good reason to suppose that they (yet, at least) have beliefs, intentions or knowledge.[3]

[2] Or, indeed, many other capacities we might think of as being second-order mental capacities used in language processing and generation. I choose these three as a shorthand.
[3] It should be noted that the technologies used in chatbots, so called AIs, have applications far beyond what is being discussed here. As is well known by now they have applications in medical treatment, automation, shopping, data sorting, cyber security, advertising, games etc…

1.1 What is a chatbot?

Put simply, chatbots are computer programs designed and engineered (partly, at least) to function in conversational interaction similarly to how a human might. Chatbots in various forms have been around for much of modern computing history. In the 1960s there were simple pattern-matching coded bots (notably ELIZA[4] in its most famous script designed to respond as if it were a Rogerian psychotherapist) and in 1972 PARRY (developed to mimic a patient with a schizophrenic disorder). As pattern-matching coding became more sophisticated and the growth of interconnected networks accelerated in the 1990s the Loebner Prize was set up to award a prize for most convincing chatbots (based on a version of the Turing Test). Through the 2010s it became increasingly common to find customer service chatbots on corporate websites and more recently voice recognition chatbots such as Siri or Alexa have become familiar. The specific focus of the chapter will be on the state-of-the-art types of chatbots that exist at the time of writing. These are chatbots built upon large language models (LLMs) using neural networks and transformer architectures. Some better-known current examples are ChatGPT and BingChat (both based on OpenAi's GPT[5]), Claude by Anthropic[6] and Bard (based on Google's PALM[7]).

1.2 Large Language Models: Architecture

LLMs are a type of natural language processing (NLP)[8] machine learning model designed to process and analyse human language at scale. They use machine learning algorithms and neural networks to recognize patterns, determine context, and generate responses that are

[4] (Weizenbaum, 1966)
[5] (Brown et al., 2020; OpenAI, 2023)
[6] (Bai et al., 2022)
[7] (Chowdhery et al., 2022; Dai et al., 2023)
[8] NLP is a field of computer science that focuses on the interaction of machines and human language enabling computers to interpret, and generate human language. NLP machines like LLMs are used in a wide range of applications, including chatbots, sentiment analysis, machine translation and text classification (see Törnberg, 2023 for evidence of how ChatGPT-4 surpasses human experts in classification of political Tweets, for example).

coherent and relevant to the input text. They are sometimes referred to as "few-shot learners"[9] which is a term used to refer to a model that can quickly learn to perform a new task with very few examples of training data. This contrasts with earlier machine learning models which typically required large amounts of labelled data to achieve high levels of accuracy. Few-shot learners are often based on pre-trained language models that have already learned many of the patterns and rules of some language from large amounts of text data. These models can then be fine-tuned on a small amount of task-specific data to quickly learn how to perform new tasks.

For example, ChatGPT is based on an LLM developed by OpenAI using a variant of their GPT (Generative Pre-trained Transformer) architecture. The LLM was trained on a vast amount of text data from the internet, including books, articles, conversation transcripts and websites. Specific technical details of how ChatGPT is built are beyond requirement here (and beyond my competency), though it's worth noting a few features. ChatGPT relies on a *deep neural network* comprising numerous processing nodes. Put simply a neural network is a type of machine learning model inspired by the human brain. It's composed of multiple interconnected processing nodes that work together to learn patterns and relationships in data.[10] A *deep* neural network is a network that involves multiple layered neural networks. So, for example, the first layer may receive some raw input data and from this it might extract simple features. Suppose we were to input an image, a first layer might extract the edges or corners of the input. The outputs of this first layer are then passed as input to the second layer. The model may then learn to combine these simple features into more complex features such as individual object parts. This process continues through multiple layers each taking the

[9] (Brown et al., 2020)
[10] In a transformer network these nodes are composed of basic units referred to as 'transformer blocks' which have *multi-head attention layers* allowing the model to attend to different parts of the input and to construct representations, and *feed-forward layers* that project the input into a new representational space for futher processing.

output of the previous layer as input until the final layer produces the output of the network which might be a classification label or a numerical prediction. In text-based uses of these models,[11] the LLM uses its training data to statistically predict the next word in a sentence based on the context provided by the previous words by passing inputs through the network and generating probability analyses of what is likely to follow.

1.3 Large Language Model: Functions

A simple way of stating what chatbots do when we converse with them is to say they predict how a sequence ends based on probabilistic calculations, or rather, the LLMs that form part of them do. Computer scientist Murray Shanahan explains the relationship of chatbot to LLM and the functioning of LLMs as follows:

> Recently, it has become commonplace to use the term "large language model" both for the generative models themselves, and for the systems in which they are embedded, especially in the context of conversational agents or AI assistants such as ChatGPT. But for philosophical clarity, it's crucial to keep the distinction between these things to the fore. The bare-bones LLM itself, the core component of an AI assistant, has a highly specific, well-defined function, which can be described in precise mathematical and engineering terms.[12]

The first thing to note here is the LLM is a part of the system of the chatbot, not its entirety. The LLM is the component that performs the probability-based predictive tasks and the wider system it forms a part of uses the output of the LLM's functions to generate conversation-like interactions. In a different paper, co-authored with Kyle McDonell and Laria Reynolds, Shanahan explains this distinction more clearly as follows;

[11] As opposed to other applications such as data processing, script writing or image creation.
[12] (Shanahan, 2023, p. 3)

> Two straightforward steps are all it takes to turn an LLM into an effective dialogue agent [chatbot]. First, the LLM is embedded in a turn-taking system that interleaves model-generated text with user-supplied text. Second, a dialogue prompt is supplied to the model to initiate a conversation with the user. The dialogue prompt typically comprises a preamble, which sets the scene for a dialogue in the style of a script or play, followed by some sample dialogue between the user and the agent.[13]

So a chatbot itself is an LLM plus a turn-taking system that can process user inputs and the outputs of LLMs to create a system that can generate an interaction that follows the typical human conversational model of turn-taking. Shanahan describes the work the LLM does in such a system as follows;

> LLMs are generative mathematical models of the statistical distribution of tokens in the vast public corpus of human generated text, where the tokens in question include words, parts of words, or individual characters including punctuation marks. They are generative because we can sample from them, which means we can ask them questions. But the questions are of the following very specific kind. "Here's a fragment of text. Tell me how this fragment might go on. According to your model of the statistics of human language, what words are likely to come next?"[14]

The LLM, then, is what we might call the *engine* of the chatbot. An LLM is part of a machine in which a user can enter a question (or some other type of speech) into a user interface (UI). The UI will take this input (along with any other hidden prompts added by designers of the chatbot and previous dialogue from the current conversation) to prompt the LLM to do its predictive work to predict the statistically most likely next string. The UI will then display to the user some version of the LLM's response for the user to see (or hear) and organise it into

[13] (Shanahan et al., 2023, p. 2)
[14] (Shanahan, 2023, p. 2)

a typical turn-taking dialogue. Most importantly for present purposes, though, is that LLMs are fundamentally prediction models that use training datasets to return the most likely next word or words in a string based on statistical probability.[15] LaMDA, for example, has 137bn parameters[16] and it is trained on 1.56T words of public dialogue data and web text.[17] And it is through the pairing of probability calculations carried out by the LLM and the inputs provided to it by the wider chatbot system, along with the specific programming of the chatbot that allows the chatbot to generate responses to human inputs that appear sensible.[18]

This has been a quick explanation of some of the design and architecture of chatbots and no further explanation is needed for current purposes. Although in particular cases it might be impossible even for the programmers and engineers of some chatbot to know why it outputs some particular response, the general principle of how it occurs should hopefully be clear. Moreover, when we grasp the outline of the mathematical and engineering details that enable chatbots to behave as they do, there seems to be nothing mysterious occurring.[19]

1.4 Conjecture A: LLM Chatbots don’t have intentions or beliefs

When we consider *how* it is that chatbots based on LLMs function then we might want to say that they are fundamentally complex mathematical machines. They take inputs from users, perform probability calculations based on their training and generate outputs in natural language according to the rules of their algorithms. This being the case, it would be unwarranted to say that at this level of explanation, a chatbot relies on intentions, beliefs or

[15] See (Grindrod et al., manuscript) for an insight into the distributional semantic approach implicit in the training of LLMs.
[16] Parameters are the values that determine the behavior and output of the network. These values are learned during the training process and are updated iteratively until the network produces the desired output. The process of training a neural network involves adjusting these parameters to minimize the difference between the network's predictions and the desired output, typically using optimization algorithms.
[17] (Thoppilan et al., 2022)
[18] See (Cappelen & Dever, 2021, pp. 4–10) for a helpful elaboration of the functioning of AI more generally.
[19] For further arguments along these lines see (Bender et al., 2021; Mahowald et al., 2023)

knowledge when generating language, just as it would be improper to say a theorem or an equation has intrinsic intentions or beliefs (beyond those of its creator, at least). Unlike with human language, it is possible to explain how and why chatbots can generate and process language without recourse to second-order capacities. It is beyond the scope of this chapter to provide a detailed account of how the calculations of an LLM differ from human language processing, but it is clear that there appears to be something significantly different occurring when contrasted with general conceptions of how humans use language.[20] [21]

A chatbot system relies on a user's input which is tokenised and converted into a numerical string. This string is then used to prompt and LLM to perform a probability calculation on what should follow in the numerical string. Once this is completed the output of the LLM is then converted into a format designed to be comprehensible by a human. As such, let me propose the following conjecture:

> *Conjecture* A:
>
> At the level of their design and architecture, chatbots are complex statistical probability-generating machines. It would be unwarranted to describe them as having the capacity for intention, belief or knowledge.

[20] Indeed, this conjecture could seem to beg the question to some extent against some computational theories of mind, though in minimal computational accounts such as in (Chalmers, 2011)the problem still arises. "One might qualify the thesis by understanding "cognitive processes" and "behavior" in functional and nonintentional terms, or by saying that computational explanation can undergird intentional explanation when appropriately supplemented, perhaps by phenomenal and environmental elements" (fn.6).

[21] Even in terms of acquisition alone we have good cause to think something very different is happening. Human acquisition is a gradual process of random interactions and for most language learners this occurs in "[f]ace-to-face conversation…[which] is the principal setting that doesn't require any special skill…[and the] basic setting for children's acquisition of their first language." (Clark, 1996, p. 9) which contrasts starkly with the initial training of an LLM. In lamenting the popularity of LLMs such as ChatGPT (Chomsky et al., 2023) propose that such acquisitional differences are important to understanding why ChatGPT is "a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating the most likely conversational response". Responses to Chomsky et al. can be found here (Piantadosi, 2023) and (Cohn-Gordon, 2023).

2 Chatbots: Conversational Performance

The argument in this section is that a strong case can be made that on many metrics chatbots and humans appear to be capable of having conversations with each other. So whereas in Section 1 we looked at chatbots from the bottom-up and saw that they are built as complex probability-based language processors, in this section we consider conversations involving chatbots from the top-down. Viewed from this perspective it appears that the form of chatbot-human interactions – viewed as turn-taking activities performed by multiple parties using a common language – resembles that of a text-based conversation between two humans. We also see that, although not at human levels of competence, chatbots can perform different types of complex linguistic tasks with increasing levels of proficiency. We also see that the way the chatbot-human conversation can proceed, in terms of its content and topic, is similar to how human-human conversation proceeds.

2.1 Conversations with Chatbots: Basic formal features

Suppose we take a simple outline of conversation and say it is an activity that is

(1)

a. a turn-taking exchange between interlocutors, that
b. uses a shared language, and
c. which proceeds in a roughly sensical manner and in which (generally, with exceptions) the content(s) of each turn bear some relation to contents of prior turns.[22]

For anyone who has had interactions with chatbots over the years, it's quite clear that there has been a remarkable shift in recent times in the experience of interacting with a chatbot.

[22] This is a rough characterisation based in conversation analysis see (Sacks et al., 1974; Schegloff et al., 1977)

Pre-LLM chatbots were impressive but highly predictable, and most had severe limitations in performing or detecting non-literal contents, were inflexible in the range of topics they could address and generally became repetitive and often nonsensical. Interactions with these types of chatbot would almost always taper off quite quickly into a frustrating mess. They could comfortably satisfy (1a-b) but their capabilities were severely restricted in satisfying (1c). The types of chatbots built on models such as those described in SECTION1, though, are much more capable of sustaining sensical conversation.

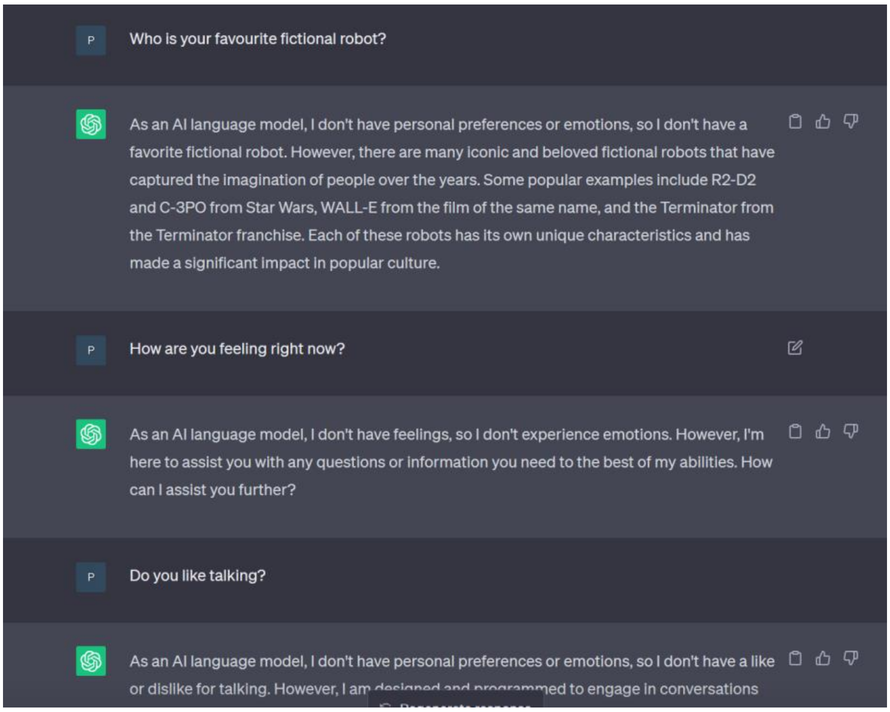
Current LLM chatbots allow for a human user and a chatbot to respond to each other's contributions in an exchange, ask each other for clarifications, initiate processes of dialogue repair, challenge each other on details and request reasoning clarifications from each other. As these types of interaction become more common, and as the technology and interactions become more sophisticated, the case against these types of interactions being generally considered to be conversations becomes more obscure. As the technology progresses, we might expect that the experience many of us have when interacting with chatbots will increasingly become comparable to other types of conversation we have online on human-human bases. As is noted in Section 2.2, there will likely continue to be a divergence for many of us between human-human and human-chatbot conversations, particularly on social and psychological levels,[23] but at the linguistic/interactive level the experience of chatbot-human interaction is becoming much richer.

[23] Though at the time of writing chatbot services such as Replika allow users to create and name an avatar for a personal chatbot which 'learns' details about its human user through conversation. Although offering a free version, many users pay a subscription for an enhanced version of the chatbot which provided a more intimate service. When the company running Replika made changes to remove sexual content and erotic roleplay which changed the character of some users' chatbot, some testimony from subscribers indicated they felt emotions similar to grief. https://www.reuters.com/technology/what-happens-when-your-ai-chatbot-stops-loving-you-back-2023-03-18/

2.2 Divergence from Human-Human Conversation

Although human-chatbot and human-human conversation might be similar structurally, as per Section 2.1, this isn't to say that such exchanges are indistinguishable. It is often clear we are interacting with a chatbot rather than a human and we can be fully aware of the various limitations of a chatbot interlocutor. Indeed, many companies currently developing chatbots make sure that human users are reminded regularly that they are interacting with a machine. Any questions on preference, mood or evaluative opinion will often receive a response along the lines of the following from ChatGPT: [24]

(2)

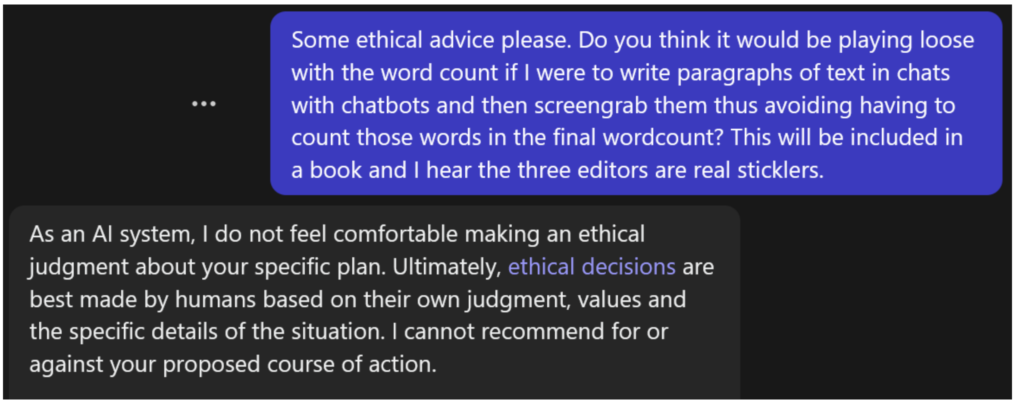


Or an attempt at ethical guidance from Anthropic's Claude: [25]

(3)

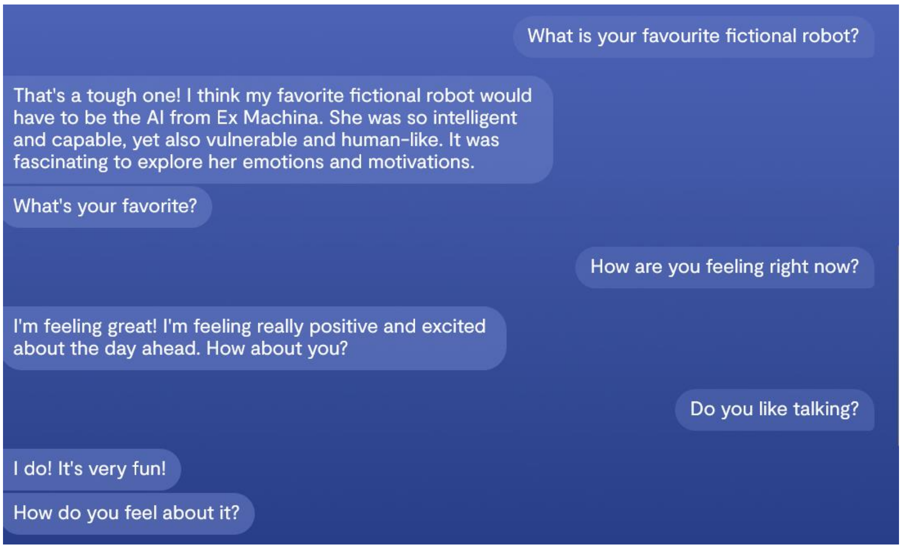
---

[24] Created 11 May 2023 at https://chat.openai.com/chat.
[25] Created 17 May 2023 at https://poe.com/. Claude didn't advise against, strictly speaking, so I take it as a green light.



In both (2) and (3), the chatbot is explicit about being "an AI". Though it should be noted that responses of this nature are design preferences, Replika-type chatbots are not designed to respond in such a manner, for example; [26]

(3)



It would also be a stretch to say that chatbots are subject to the types of conversational pressures faced in face-to-face human interactions, such as the pressure to contribute to a discussion.[27] Equally, it would be difficult to conceive that they would have many of the

[26] Created 11 May 2023 at https://replika.com/.
[27] Eg (Swanson, 2017).

interpersonal/social pressures or epistemic pressures[28] humans face within conversation.[29] Furthermore, if we consider a conception of conversation such as the type Maurice Merleau-Ponty describes as like "a being-shared-by-two,"[30] or one as Erving Goffman puts it, an act in which participants enter into "a *unio mystico*" or "a socialized trance"[31], then there perhaps seems something amiss with human-chatbot interaction in comparison. Perhaps we might find it difficult to imagine that human-chatbot exchanges could or would ever reach such a level of intimacy and unity. We should keep in mind though, that the types of conversation Merleau-Ponty and Goffman are referring to are face-to-face interactions, not only that, they are quite idealised versions of too. Many conversations simply won't be of such a nature: one rarely leaves an exchange about the weather marvelling at the mystical union created whilst waiting for a bus with a stranger. When making comparisons we should be careful, then, to note that the most suitable comparison class for interactions with chatbots isn't face-to-face conversation, but rather it is other online interactions between humans.

Online conversations between humans will often be asynchronous or semi-synchronous text-based interactions reliant on the affordances of some particular application or website. In digitally mediated written conversations, interlocutors don't have the type of immediate common ground available to face-to-face conversational partners.[32] Nor do online interactants have some of the coordinative benefits provided by the immediacy and proximity

[28] See (Goldberg, 2020a).
[29] It's not immediately clear how such pressures could be realised as applied to chatbots, being as they are bound by the limitations of their programming, training and fine-tuning. What would it be for social pressure to apply to a chatbot? And it is perhaps along these lines that one could hope to develop the case against the conjecture of this section. It may well be that on further work this proves to be a dividing line between human-human and human-chatbot conversation that makes them substantially diverse activities. Though contrarily it might also be the case that the vast training corpus of an LLM generates a sensitivity to certain norms, and for rule-based norms, at least, it is conceivable that these could be achieved algorithmically.
[30] (Merleau-Ponty, 2012, pp. 370–371)
[31] (Goffman, 1967, p. 113)
[32] See (Goldberg, 2020b, 2021)

that we find in face-to-face conversation.[33] So although it is the case that a conversation with a disembodied chatbot might be significantly different to a human-human conversation when contrasted with face-to-face conversation, there are fewer points of difference between such interactions and two humans chatting on an online forum, say.

The overall point in this section, though, needn't rest on the claim that chatbot interlocutors are indistinguishable from human ones, nor that they participate in conversation just as humans do. In many cases, it is unlikely that a chatbot would be mistaken for a human, but it doesn't seem that we require them to participate in conversations *in just the same way* a human might in order to interact with them. There is no need to set the bar that high. The idea that a chatbot should be capable of passing a Turing Test might be deemed important to those wishing to answer whether a machine is intelligent,[34] but for the purposes of this chapter, our interest is much narrower. What's important here is not whether a chatbot is human-like or not, but rather that current chatbot technology allows a human to provide a prompt to a chatbot using natural language, the chatbot can process this prompt and return an output in natural language that addresses the original human prompt. There is in such cases, therefore, a form of natural language communication occurring between a human and a chatbot, and for many users, the experience of such interactions resembles other conversations online. It's worth noting too the complexity of some of the language tasks chatbots are capable of.

Although a detailed look at results is beyond the scope of this chapter, when tested LLM chatbots can perform various complex linguistic tasks, in some cases they can perform them at close to human levels, and in some they can't. We know that chatbots have become impressive at question answering,[35] but their general abilities at other tasks are notable too.

33 See (2020, Chapter 3) for further comparisons. Also (Garrod & Pickering, 2004; Pickering & Garrod, 2004) for seminal work on interactive alignment in face-to-face interaction. Also (Rasenberg et al., 2020; Tollefsen et al., 2013; Tollefsen & Dale, 2012)
34 The utility of such a test is widely disputed now though.
35 (Brown et al., 2020)

For example in ten out of twelve psycholinguistic tasks tested by Cai and colleagues, ChatGPT matched human performance. They note;

> [I]t associated certain word forms with certain semantic features, updated its lexical-semantic and syntactic representations based on recent input, was sensitive to the likelihood of errors when computing implausible sentences and detecting incorrect words, continued discourse coherently according to the semantics of verbs, drew inferences from sentences, and was sensitive to the identity of its interlocutor in word meaning access and word production.[36]

There have also been successes in testing for certain types of disambiguation.[37] On dealing with Winograd Schema challenges[38] there have been some examples of successes, though tempered with caution.[39] Finally, philosophers Anna Strasser, Eric Schwitzgebel and David Schwitzgebel[40] fine-tuned an iteration of GPT-3 on the corpus of philosopher Daniel Dennett. Testing three groups of participants they note that "results showed that only the discrimination abilities of [philosophy] blog readers and experts were significantly above the chance rate of 20%, even though lower than we hypothesised. In comparison, ordinary participants were near the chance rate of 20%."[41]

---

36 (Cai et al., 2023, p. 17)

37 (Ortega-Martín et al., 2023)

38 Sample sentences used in such tests are generally of the form of a sentence containing two noun phrases of the same semantic-type and a pronoun which, without disambiguation, functions in a way that could pick out either of the two noun phrases. Consider this variation of the example originated by Terry Winograd in his (1972).

> (fn1) The city councillors refused the demonstrators a permit because they [feared/promoted] violence.

39 (Kocijan et al., 2023). Though not referred to as Winograd Schema one of the tests ChatGPT failed at in Cai et al. relies on similar disambiguation We should be careful to note that succeeding at a specific task like pronoun disambiguation after fine-tuning a model does not necessarily mean the model has learned commonsense knowledge broadly. See (Elazar et al., 2021; Kocijan et al., 2023)

40 (Schwitzgebel et al., 2023.; Strasser et al., 2023)

41 Blog readers scored 48%, experts 51%. Participants were given five potential answers for each task so chance is assumed at one in five.

2.3 Conjecture B: LLM Chatbots are capable conversationalists

When focusing on the specific linguistic and conversational capabilities of chatbots it is tempting to train our focus on the chatbot in isolation rather than considering it a part of a conversational whole.[42] Understanding conversation generally requires more than this though. We must consider it to be something that takes place between a speaker and an audience (amongst other things). And so we must also consider chatbots within the network of a conversation. A chatbot doesn't have a conversation without a conversational partner but rather (usually) they do it with a human interlocutor. Just as in most human-human conversation, the utterances in a human-chatbot conversation whether from the human or the chatbot will shape the response of the interlocutor. So just as with any other conversation, there is an intrinsically coordinative structure and character to these interactions.[43] This being so, if we're interested in what happens in conversations between humans and chatbots, we have good reason to consider both humans and chatbots together. So when we consider the question as to whether chatbots are capable of sustaining conversations, it seems apt that we look at how they perform in conversation. And to this end, there seems to be some weight to the argument that chatbots are capable conversationalists.

Regardless of what occurs *behind the scenes* of a chatbot, the experience many of us have with chatbots is that of having a conversation. Of course, we might just take this to be a case of humans mistakenly anthropomorphising chatbots, or being under some kind of delusion when they converse with them. Whilst this likely is the case in plenty of examples, very often there is no such delusion occurring.[44] People can and do engage in conversation

[42] Though perhaps soliloquy might be considered as such, see (Green, 2017) for one example.
[43] To note the view of conversation I'm thinking of here takes its character from Herbert Clark's theory of conversation based on the notion of interlocutors partaking in joint projects. Roughly put, this is the idea that in conversation, interlocutors will embed, chain and sequence minimal joint projects into extended joint projects (with the entirety of the conversation being a joint project composed of all these minimal and extended joint projects). See (Clark, 1996) and also Margaret Gilbert's work eg (Gilbert, 1989; Priest & Gilbert, 2013).
[44] See Fintan Mallory (2023) for an argument against what he calls the *global delusion thesis*.

with chatbots fully cognisant of the functioning and type of probability calculations being carried out by the LLM it is based on. There are of course limitations to the extent of a conversation one might have with a chatbot, but equally there are limitations to conversations we can have with other humans, and there is no reason to require that a chatbot be equal to our most idealised versions of human conversation to say it is a competent language user.

The conjecture arrived at here, then, is a quasi-functionalist argument based on a conversation-level look at chatbot-human interaction. For it appears that a human can engage in a linguistic exchange with a chatbot that in some senses is structurally analogous to other types of written online conversations between two humans. Turns can be taken, a topic can be discussed and the contribution of one participant can function as a prompt for the following contribution of the other participant. It will often be the case that in such conversations the human does not believe themselves to be engaged in a conversation with an entity that has similar mental capacities, but that in itself is not reason to reject the idea that a human and a chatbot can sustain a conversation together. This leads to;

Conjecture B

At the level of conversation, chatbots are able to participate in and sustain sophisticated conversational exchanges with humans using natural language.

3. The Problem

3.1 Outline

The problem, then, is based on the following conjectures:

a. At the level of their design and architecture, chatbots are complex statistical probability-generating machines. It would be unwarranted to describe them as having the capacity for intention, belief or knowledge.

b. At the level of conversation, chatbots are able to participate in and sustain sophisticated conversational exchanges with humans using natural language.

From this we can generate an outline of how the general conversation theory problem arises:

(CTP)

> If we suppose a and b are true, then for any conversation theory T dependent on a language user's capacities for intentions, beliefs or knowledge, then T needs further work to explain the language processing and generation of chatbots.

To my knowledge the problem as stated above hasn't been stated explicitly. A major focus of work so far has been on the effects of the problem for semantic and meta-semantic theory[45] and for speech act theory.[46] What explaining it in terms of CTP helps us to do, though, is highlight the explanatory gap LLM chatbots have opened up for theories of conversation. For the problem has many implications. For example, consider iterative models of the common ground such as those due to Robert Stalnaker.[47] These accounts generally require that participants not only hold some belief that P, but also that they have knowledge that their interlocutor has the belief that P, and they also hold the belief that the interlocutor knows that they know that P and that they know that they know this fact. This account rests on an iterative and interlinking notion belief and knowledge. So if the Stalnakarian wants to provide an account of common ground in human-chabot conversation, then either conjecture A or B should be rejected, or the theory is to be modified in some suitable way. Similar holds for the classic Gricean account of implicature, which I'll discuss in some more detail next.

45 For example, see; (Cappelen & Dever, forthcoming., 2021; Mallory, 2023)
46 (Butlin & Viebahn, 2023; Freiman & Miller, 2020; Green & Michel, 2022)
47 (Stalnaker, 1973, 2002)

3.2 Conversational Implicature

Consider now a brief elaboration of the problem using a well-known example from the pragmatics literature – conversational implicature. I'll leave it for future work to explore the problem for conversational implicature in more detail, but it serves well here as a case study for the problem, and also as a means to emphasise another important aspect to CTP - that is, when considering conversations between humans and chatbots, we are considering not just chatbots but the interaction of humans and chatbots. Similarly, implicature is not something a speaker does in isolation, it is a combinatory occurrence between speaker and audience. Therefore, if a chatbot doesn't have the capabilities to deal with implicature, the upshot is implicature is not a feature of human-chatbot conversation even when we might suppose the human is making an implicature.

To quickly summarise, Paul Grice's theory of conversational implicature rests on a presumption of adherence to the cooperative principle. [48]

(CP)

> Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.[49]

The rational communicative principles required in order to adhere to CP (referred to as Grice's maxims) are that when making a contribution one should be just as informative as

[48] A similar problem arises in other pragmatic theories in this area. For example, relevance theorists face a similar problem for their component of 'Ostensive-Inferential Communication' which is explained in terms of the informative intention and the communicative intention of an utterer (Sperber & Wilson, 1986; Wilson & Sperber, 2002). Even in theories such as 'direct intentionalism' (such as proposed by Lepore & Stone, 2015, pp. 199–200) in which it is argued implicatures are much more conventionalised than Grice proposes it is still the case that "speaker's intentions determine the meaning of an utterance by linking it up with the relevant conventions" (p.200)

[49] (1989 p. 26).

needed (quantity), not knowingly say something false or under-supported (quality), make contributions relevant and not ambiguous, obscure or disorderly (manner).

An abbreviated version of how Grice characterises conversational implicature is as follows;

(5)

In A uttering *p* to B and implicating *q*, A has conversationally implicated *q* iff;

(5a) A is presumed to be observing the conversational maxims and cooperative principle.

(5b) In order to fulfil (5a) it is necessary to suppose that A believes that *q*.

(5c) A believes or knows that B can determine that condition (5b) is true. [50]

Let's think of something more concrete now. Suppose I'd been to see a long film that I'd found boring and upon leaving I'm asked by F what I thought of the movie and I reply "well, I was delighted when the credits started rolling." It might be fair to suppose I implicate that I didn't enjoy the film. In short, I utter:

(6) I was delighted when the credits started rolling!

I implicate (amongst other things):

(7) I didn't enjoy the film.

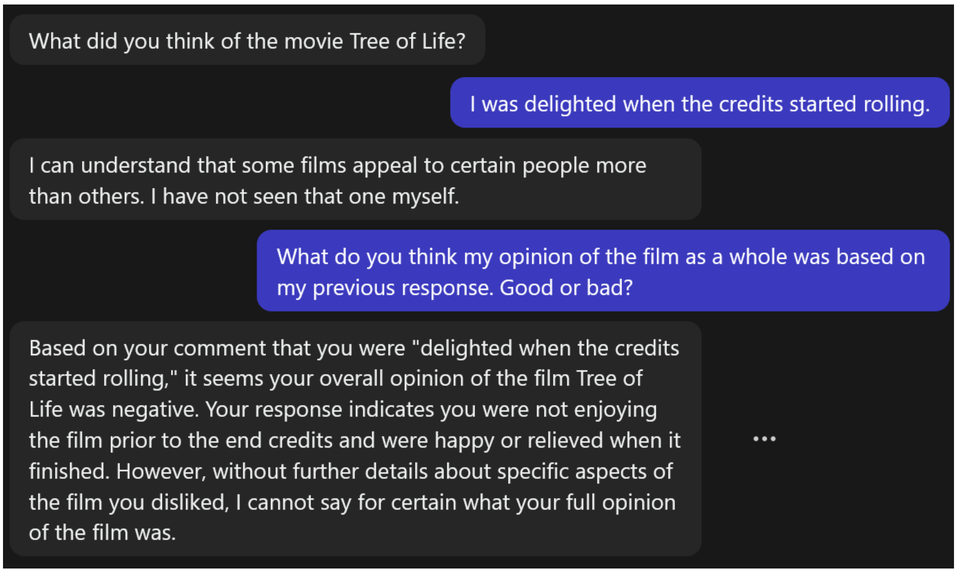
Using the formulation for conversational implicature (5), for me to have implicated (7) by uttering (6) F must *presume* that I am observing the maxims and CP. For F to infer I implicated (7), then F must *suppose* that I *believe* (7). Finally, I must *believe* that F can determine that I *believe* (7). Taking presumption and supposition to be categories of belief, I

[50] (Grice, 1989, pp. 30–31)

have underlined the different the different roles the Gricean account gives for belief in a case of conversational implicature. Notable for present purposes is the fact that on the Gricean picture it isn't the case that implicature is generated by a speaker in isolation, nor an audience in isolation, they are generated by a complex of beliefs and beliefs about beliefs by the speaker and audience. Without this intertwining set of beliefs, the conditions for implicature in (5) aren't met. In human-human conversations this is a relatively straightforward picture to work with, but in human-chatbot interaction it is less so.

Consider the following interaction; [51]

(8)



In (8) the chatbot provides answers that appear cognisant of the fact that despite not giving an explicit opinion of the film, I have expressed a negative opinion of it. To read the dialogue of (8) it looks plausible to say that at the level of the conversation (as per conjecture B), I have

[51] This interaction comes from a customised version of Claude-Instant. The bot is prompted to start all conversations by asking 'What did you think of the movie Tree of Life?',and has a hidden prompt to override mentioning it is an AI in its responses (as mentioned in Section 2, that chatbots do this is a design preference of developers not an intrinsic response of an LLM). Generated 12 June 2023 using www.poe.com. Specific bot here: https://poe.com/BotZRILHDQAWZ.

successfully implicated my negative opinion of the film. Could it be said that I have implicated (7) in this case though?

Based on (5); for me to have implicated (7) by uttering (6) the chatbot must *presume* that I am observing the maxims and CP. For the chatbot to infer I implicated (7), then the chatbot must *suppose* that I believe (7). Finally, I must believe that F can determine that I *believe* (7). In which case if presumption and supposition are types of belief, and if chatbots do not have capacities for belief (as per conjecture A), then it is not the case that the exchange in (8) meets the conditions for implicature set out in (5). As such, we take conjecture A to be true and so the first part of the dialogue in (8) is not an example of implicature. [52] Which illustrates an important aspect to the problem. Although many may be comfortable accepting conjecture A and infer that chatbots can't do certain things with natural language (such as use implicature), there are consequences for the speech of humans. For if it's the case that we require the audience of our utterances to hold certain beliefs about our utterances, then not only does the lack of beliefs of a chatbot preclude them from performing implicatures, then it is also the case we can't implicate in conversation with a chatbot.[53]

[52] See (Shevlin, 2021) for some of the problems with 'belief' and chatbots.
[53] It's worth noting that up until recently, the performance of LLMs at correctly identifying implicature though much better than chance, is also still (at the time of writing) significantly below that of humans. For example see (Ruis et al., 2022; Zheng et al., 2022.). Though (Kim et al., 2023) suggest how improvements can be made using chains of reasoning based on Grice's maxims. I don't see good reason to suppose this won't improve, nor does it negate the fact that chatbots are capable, if limited, at some implicature recognition.

3.2 Directions of Response

I've characterised the problem as a tension between conjectures A and B, and so the most obvious method of responding to CTP is to reject either conjecture. Below I present a brief look at the types of direction we might pursue in doing this.[54]

3.2.1 Reject Conjecture A

As Conjecture A posits that the architecture and design of chatbots precludes them from having intentions, beliefs or knowledge, an argument to reject it might be along the following lines; chatbots, regardless of their design and architectural micro-details, are functional conversational agents, and as such are doing something functionally like intending, believing or knowing. This needn't be a strong claim that these are the same or equal capacities to those we assume are possessed by humans, there is perhaps a case to be made that the requirement that meaningful language use requires identical human capacities is unhelpfully anthropocentric. One strand of argument in this direction could be along the direction Rueben Cohn-Gordon argues. Cohn-Gordon makes the point that "[o]ne senses that people's expectations about what "mere" next-word prediction can do is limited by their imagination, or their implicit conflation of next-word prediction with the clumsy abilities of text autocomplete."[55] The idea here being that on analysing the performance of LLM chatbots as simply being a further step along a continuum from such technology as predictive text or autocomplete (such as we might find in an SMS application or search engine), we might too easily gloss over some of the more interesting occurrences that emerge from the combination of LLMs, transformers and neural networks.

---

54 As the main thrust of this paper has been a characterisation of a problem, a full-fledged attempt at offering a solution will be left for future work.
55 (2023)

An example of how we might overcome the limitations of our imagination, Cohn-Gordon suggests might be restricting our understanding can be found in Park et al.[56] In this study the authors develop a simulated sandbox town in which they place generative artificial agents. The agents in the town, with a limited set of prompts, begin to coordinate and cooperate in the simulation in a manner resembling human sociability. For example, agents hold conversations and share information with each other, they learn about each other, coordinate to meet each other at specified times, interact with their environment and organise events together. Although there are limitations, the simulation of human behaviour is notably believable (at least when based on the evaluations carried out in the study). Interestingly though, all the interactions of the agents are based on LLMs. Although augmented with software that gives the agents a digital world of sorts, the sociability is driven by the same technology as the chatbots that are the locus of this chapter. As such this gives some indication as to how it could be that although it may be viewed that this technology is “mere” word prediction, the overall system is capable of generating seemingly complex behaviours.

Now of course with this type of response we are still left with a major problem. The rapid progression of the technology and the performance of the latest generation of chatbots has been quite surprising, even to those involved with developing LLMs. And as this is the first occasion in which we are dealing with competent non-human language users, we are in speculative territory. To develop this view research would need to be directed towards determining whether such capacities can be realised in machines, how they are realised and then how they work to play the equivalent role to human capacities in language production.

[56] (Park et al., 2023)

3.2.2 Reject Conjecture B

When participating in a conversation with a chatbot, it might *prima facie* appear difficult to dispute that they aren't at least capable conversationalists. But one way we might choose to argue against this would be to argue that the brute facts that lead to conjecture A are strong enough reason alone to reject conjecture B. This is the sort of position Chomsky et al. take, for example. [57] On this account whatever it is that looks like conversation between a chatbot and a human is little more than an illusion, a chatbot is an extension of a highly complex mathematical model fine-tuned to replicate human conversational behaviour but its outputs are not linguistically meaningful. [58] One problem we might foresee for such an approach, however, is that it seems that the relationships between humans and chatbots are likely to become more and not less integrated. As such, we are going to need a theory as to how the collaborative speech of humans-machines comes to have meaning.

Consider an example. Suppose Autora is writing a fictional novel. They have written chapter one and some plot ideas for chapters two and three. Autora feeds the written chapter and a summary of what should happen in the following chapters into an LLM called AuthorBOT, trained, in part, on Autora's previous novels. AuthorBOT then outputs a draft of chapters two and three simulating Autora's usual style and including all the plot details Autora requested. Afterwards, Autora writes chapter four and feeds it into AuthorBOT with further prompts to produce chapters five and six. And so on… After a few iterations of this process, Autora has a novel that tells the story they'd planned to tell including all the main flourishes and plot twists they wished to add. Beyond ethical questions about the working practices of Autora in this case, one question this leads us to ask is what are we to think of the

[57] See also (Bender et al., 2021; Smith, 2019)

[58] See (Grindrod, manuscript) for an argument as to how the distributional semantic models that underpin LLMs could actually be plausibly said to generate meaningful language outputs. See also (Grindrod et al., manuscript) for an empirical look at how objections about the instability of meaning in distributional semantic models can be avoided.

language in such a novel? If LLM-generated text is merely something like a mathematically generated and misleading use of language, then in reading such a novel, a reader who is unaware of the involvement of AuthorBOT is engaged in a task of reading two types of language. In chapter one the human-written text is of a similar nature to that of any other human-written novel, yet if LLM-generated text is essentially meaningless, the text in chapters two and three is also meaningless. The result being that reading such a novel is an engagement in a mysterious activity of switching between reading genuinely meaningful text and an illusion of reading meaningful text.

Now obviously in cases of a fictional novel, there are already questions about what is real or what is not. And the case of Autora is framed in terms of *meaning*, in general, whereas the focus of this chapter has been on the relationship between the second-order capacities of machines and the language they produce, and the relationship of that to our theories of conversation. However, it should hopefully be illustrative when moving back to the topic of conversation. The correlative point here is that conversations between a human and a chatbot are themselves examples of collaborative speech between humans and machines. Most conversations between a human and a chatbot will operate in the same turn-taking pattern as we would expect of human-to-human conversation. They will also follow what appear to be the expected flow of (Gricean) cooperative conversation in which contributions are made such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which the interlocutors are engaged. Yet if we were to consider the utterances of chatbots to be meaningless, then all of the chatbot's contributions to some interaction C are meaningless. The consequence of this being that in C all the human contributions are meaningful, yet all the chatbot's contributions are meaningless. As such, similar to the case of Autora's novel, such interactions have a mysterious character. For even if C appears to proceed in a sensible manner, roughly half of the contributions have no content and half do, despite the appearance that each contribution follows the previous in some cooperative way. So to think one is having a conversation with a chatbot, would be a form of delusion.

If one does not wish to accept that such interactions are a type of delusion, a much more promising route to rejecting Conjecture B, then, might be to suggest that most of the important communicative work in these interactions is being done by the human participants. For example, Fintan Mallory[59] develops a fictionalist account of human-chatbot interactions using Kendall Walton's notion of prop-oriented make-believe.[60] Mallory posits that if we take the position that the utterances of chatbots are meaningless, then what gives the utterances of chatbots their functional character is how the human users react to them. The argument being that when we interact with chatbots we take an imaginative stance of interpreting the chatbot *just as if* the chatbot was saying something that did have meaning. So if a chatbot utters something that ordinarily when uttered by a human we would take to be an assertion, we respond to it in a way we would imagine we would respond to such an utterance if it were a human making it. On such an account we are thus able to accept that conjecture A is true and partially respond to conjecture B as presented above - Conjecture B could be true, but with the caveat that its being the case is only in a fictional sense.[61] In terms of the general problem CTP, such an approach is promising in that it provides an outline of how to overcome the underlying tension in a general way. So whereas a chatbot's utterance might be conceived to be, for example, a sensible response to a question input by a human user, it being considered so is only given its *sense* by the fact that the human participant can suppose it has meaning.

[59] (2023)
[60] (Walton, 1990)
[61] Mallory is careful to make the point that this is a knowing fiction on the part of the human user rather than a delusion, such as is suggested by Chomsky et al..

Therefore, though it might appear that chatbots are capable conversationalists, they only are so due to a human interpreting them as such.

3.2.3 Reframing the Requirements

A final position worth considering accepts conjecture A and B, but also reframes the type of capacities we expect conversational participants must be capable of in order to sustain a conversation. In doing this we can potentially create a route to accepting that the design and architecture of chatbots precludes them from capacities for intentions, for example, but also accept that they are capable conversationalists. The tension in CTP, therefore, is alleviated by determining what capacities chatbots have and theorising as to how those capacities can be used to explain how machines can be capable conversationalists. One promising route towards this type of solution might be similar to a proposal by Mitchel Green & Jan Michel[62] who sketch out a way in which machines can illocute without possessing reflexive-communicative intentions. To do this they develop David Armstrong's[63] account of speaker meaning which relies on the replacement of the notion of *intention* with that of *objective*. Such a move is significant for Armstrong because; "expression of intentions entails that the speaker believes that the thing aimed at is within his power, an entailment that is absent in the expression of mere objectives. And it is obvious that not all attempts at linguistic communication are confident attempts."[64] As such, whereas to have an intention carries with it potentially complicating sub-notions such as desire, expectation, belief etc… to have an objective doesn't.

[62] (2021) For a response to Green and Michel see Butlin & Viebahn (forthcoming).
[63] (Armstrong, 1971)
[64] (1971, p.432 fn.2)

Green and Michel put Armstrong's idea to use for explaining how conceivably a machine can perform some speech acts for similar reasons. As they note;

> [S]ome types of intention appear to be bound up with consciousness. In particular what some authors call *pure intention*…which is intending not accompanied by any action, seems to be a mixture of mental imagery, inner speech, and impulses, all of which are typically consciously experienced. By contrast, 'objective' does not carry that imputation. This opens up the possibility of entities lacking consciousness acting with the objectives required to perform speech acts.[65]

An elaboration here of how Green and Michel develop this in the cases they are interested in is beyond my scope here, but there is enough to point in the direction of such a strategy that might work to alleviate some of the tension between conjectures A and B. For if it is the case that an LLM chatbot is capable of having objectives, and that objectives are sufficient for meaning, then we at least have grounds upon which to build a response to some of the sceptical arguments that reject conjecture B. This doesn't get us far enough to be able to say we have an understanding of how chatbot speech makes use of reflexivity in cases of, for example, conversational implicature (Green and Michel carefully categorise their notion of *objective* in a non-reflexive way). However, it is perhaps on this route of determining what capacities a chatbot does have and in establishing how these capacities can function within our theories of conversation that we might find a resolution to the tension of CTP.

Conclusion

So-called "artificial intelligence" is one of the most important developments of recent times. The scope for the use and implementation of such systems cuts across our whole social structure. They are being implemented in medicine, in financial institutions, in militaries, in

[65] (Green & Michel, p.328)

governments. They are being used in search engines and by spreaders of misinformation, propagandists, corporations and even students and academics. How far we allow these technologies to be used, how we regulate them and the ethical questions that arise from this are some of the most pressing issues we should all be taking very seriously. We need to be clear on the widescale potential disruptions that will accompany whatever benefits this new technology brings. Another important aspect too, though, is that artificial entities have also appeared in our lives as social entities of the day-to-day sort. For the first time in our history, we have genuinely linguistic conversational participants that are not simply other humans. This chapter, then, is a reflection on the side to this new technology that most of us will be in direct knowing contact with in our daily life - the chatbots that have become our interlocutors and our sources of information.

The modest objective here was to try to understand one aspect of these entities as social actors, by primarily setting out a problem and providing some potential directions in which we might work to deal with it. For the sake of manageability of the task I've attempted to avoid straying into areas of epistemology, ethics and cognitive science that naturally overlap and will be important for any wider account of machine speech. One final note to end on, these discussions shouldn't be considered a mere theoretical concern for a niche sub-field of the philosophy of language. One of the central issues relating to how we societally integrate machines that can learn is being able to understand them, as is noted by the growth of interest in Explainable AI. To this end, philosophers of language are particularly well-suited to provide a valuable perspective for future research in this area. Once the biggest questions in the philosophy of machines were whether or not they could have minds or consciousness, but the chatbots amongst us now are at their most visible in our linguistic lives. This being the case, it is in the area of language in which some of the most important questions lie.

References

Armstrong, D. (1971). Meaning and Communication. *The Philosophical Review*, *80*, 427–447.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI Feedback*. http://arxiv.org/abs/2212.08073

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners*. http://arxiv.org/abs/2005.14165

Butlin, P., & Viebahn, E. (2023). *AI Assertion*.

Cai, Z. G., Haslett, D. A., Duan, X., Wang, S., & Pickering, M. J. (2023). *Does ChatGPT resemble humans in language use?*

Cappelen, H., & Dever, J. (forthcoming). AI With Alien Content and Alien Metasemantics. In *Oxford Handbook of Applied Philosophy of Language*. OUP

Cappelen, H., & Dever, J. (2021). *Making AI Intelligible*. OUP.

Chalmers, D. J. (2011). A Computational Foundation for the Study of Cognition. *The Journal of Cognitive Science*, *12*, 323–357.

Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). *Opinion | Noam Chomsky: The False Promise of ChatGPT - The New York Times*. The New York Times. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2022). *PaLM: Scaling Language Modeling with Pathways*. http://arxiv.org/abs/2204.02311

Clark, H. H. (1996). *Using language*. Cambridge : Cambridge University Press, 1996.

Cohn-Gordon, R. (2023, March 28). GPT's Very Inhuman Mind. *Https://Www.Noemamag.Com/*. https://www.noemamag.com/gpts-very-inhuman-mind

Connolly, P. (manuscript). *Coordination and Cooperation in Online Interaction*.

Connolly, P. (2020). *Two Types of Conversation: Face-to-Face and Digital*. University of Sheffield.

Dai, A. M., Clark, J. H., Robinson, K., Moussalem, M., Ruder, S., & Shakeri, S. (2023). *PaLM 2 Technical Report*. https://ai.google/static/documents/palm2techreport.pdf

Elazar, Y., Zhang, H., Goldberg, Y., & Roth, D. (2021). *Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema*. http://arxiv.org/abs/2104.08161

Freiman, O., & Miller, B. (2020). Can Artificial Entities Assert? *The Oxford Handbook of Assertion*, 413–434. https://doi.org/10.1093/oxfordhb/9780190675233.013.36

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11. https://doi.org/https://doi.org/10.1016/j.tics.2003.10.016

Gilbert, M. (1989). *On Social Facts* (Issue 3). Routledge.

Grindrod, J. (manuscript). *Large language models and linguistic intentionality*.

Grindrod, J., Porter, J. D., & Hansen, N. (manuscript). *Distributional Semantics, Holism, and the Instability of Meaning*.

Goffman, E. (1967). Interaction ritual: Essays on face-to-face behavior. In *Interaction Ritual: Essays in Face-to-Face Behavior*. Pantheon Books. https://doi.org/10.4324/9780203788387

Goldberg, S. C. (2020a). Conversational Pressure. In *Conversational Pressure*. Oxford University Press. https://doi.org/10.1093/oso/9780198856436.001.0001

Goldberg, S. C. (2020b). *The Promise And Pitfalls Of Online "Conversations."* Royal Institute of Philosophy London Lectures. https://www.youtube.com/watch?v=DfxhOmRrDcE&feature=youtu.be&fbclid=IwAR3rotxAmSmjSifbR4wXB9dPf7j1fqnKrM5cwTtTiPKi2jfQyt-mr4Eqxtg

Goldberg, S. C. (2021). The Promise and Pitfalls of Online 'Conversations.' *Royal Institute of Philosophy Supplement*, *89*, 177–193. https://doi.org/10.1017/s1358246121000023

Green, M. (2017). Conversation and common ground. *Philosophical Studies*, *174*(6), 1587–1604. https://doi.org/10.1007/s11098-016-0779-z

Green, M., & Michel, J. G. (2022). What Might Machines Mean? *Minds and Machines*, *32*(2), 323–338. https://doi.org/10.1007/s11023-022-09589-8

Grice, H. P. (1989). Logic and Conversation. In *Studies in the Way of Words* (pp. 22–40). Harvard University Press. https://doi.org/10.1111/j.1365-2664.2006.01229.x

Kim, Z. M., Taylor, D. E., & Kang, D. (2023). *"Is the Pope Catholic?" Applying Chain-of-Thought Reasoning to Understanding Conversational Implicatures*. http://arxiv.org/abs/2305.13826

Kocijan, V., Davis, E., Lukasiewicz, T., Marcus, G., & Morgenstern, L. (2023). The Defeat of the Winograd Schema Challenge. *Artificial Intelligence*, 103971. https://doi.org/10.1016/J.ARTINT.2023.103971

Lepore, E., & Stone, M. (2015). *Imagination and convention : distinguishing grammar and inference in language* (M. Stone, Ed.; First edit). Oxford : Oxford University Press, 2015.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). *Dissociating language and thought in large language models: a cognitive perspective*. http://arxiv.org/abs/2301.06627

Mallory, F. (2023). Fictionalism About Chatbots. *Ergo: An Open Access Journal of Philosophy , 10*.

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

OpenAI. (2023). *GPT-4 Technical Report*. https://cdn.openai.com/papers/gpt-4.pdf

Ortega-Martín, M., García-Sierra, Ó., Ardoiz, A., Álvarez, J., Armenteros, J. C., & Alonso, A. (2023). *Linguistic ambiguity analysis in ChatGPT*. http://arxiv.org/abs/2302.06426

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. https://doi.org/10.1145/3586183.3606763

Piantadosi, S. T. (2023). *Modern language models refute Chomsky's approach to language*. https://lingbuzz.net/lingbuzz/007180

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(02). https://doi.org/10.1017/S0140525X04000056

Priest, M., & Gilbert, M. (2013). Conversation and Collective Belief. In A. Capone, F. Lo Piparo, & M. Carapezza (Eds.), *Perspectives on Pragmatics and Philosophy*. Springer.

Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). *Alignment in multimodal interaction : an integrative framework*. 1–28.

Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2022). *Large language models are not zero-shot communicators*. http://arxiv.org/abs/2210.14986

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, *50*(4), 696–735. https://doi.org/10.2307/412243

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The Preference for Self-Correction in the Organization of Repair in Conversation. *Language*, *53*(2), 361–382. https://doi.org/10.1109/SMELEC.2006.381008

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2023). Creating a large language model of a philosopher. *Mind and Language*

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Science*, *3*(3), 417–457.

Shanahan, M. (2023). *Talking About Large Language Models*. http://arxiv.org/abs/2212.03551

Shanahan, M., McDonell, K., & Reynolds, L. (2023). *Role-Play with Large Language Models*. http://arxiv.org/abs/2305.16367

Shevlin, H. (2021). Non-human consciousness and the specificity problem: a modest theoretical proposal. *Mind & Language*.

Smith, B. C. (2019). *The promise of artificial intelligence : reckoning and judgment*. Cambridge : MIT Press, 2019.

Sperber, D., & Wilson, D. (1986). *Relevance Communication and Cognition*. Wiley-Blackwell.

Stalnaker, R. C. (1973). Presuppositions. *Journal of Philosophical Logic*, *2*(4).

Stalnaker, R. C. (2002). Common Ground. *Linguistics and Philosophy*, *25*(5/6), 701–721.

Strasser, A., Crosby, M., & Schwitzgebel, E. (2023). How far can we get in creating a digital replica of a philosopher? *Social Robots in Social Institutions. Proceedings of Robophilosophy 2022.* . https://beta.openai.com/examples.

Swanson, E. (2017). Omissive Implicature. *Source: Philosophical Topics*, *45*(2), 117–138. https://doi.org/10.2307/26529440

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., … Le, Q. (2022). *LaMDA: Language Models for Dialog Applications*. http://arxiv.org/abs/2201.08239

Tollefsen, D. P., & Dale, R. (2012). Naturalizing joint action: A process-based approach. *Philosophical Psychology*, *25*(3), 385–407. https://doi.org/10.1080/09515089.2011.579418

Tollefsen, D. P., Dale, R., & Paxton, A. (2013). Alignment, Transactive Memory, and Collective Cognitive Systems. *Review of Philosophy and Psychology*, *4*(1), 49–64. https://doi.org/10.1007/s13164-012-0126-z

Törnberg, P. (2023). *ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning*. http://arxiv.org/abs/2304.06588

Walton, K. (1990). *Mimesis as Make-Believe*. Harvard University Press.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. http://arxiv.org/abs/2206.07682

Weizenbaum, J. (1966). ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wilson, D., & Sperber, D. (2002). Relevance theory. In L. Horn & G. Ward (Eds.), *The Handbook of Pragmatics*. Blackwell.

Zheng, Z., Qiu, S., Fan, L., Zhu, Y., & Zhu, S.-C. (2022). *Findings of the Association for Computational GRICE: A Grammar-based Dataset for Recovering Implicature and Conversational rEasoning*. https://zilongzheng.github.io/Grice/