Abstract

According to widely accepted views in metasemantics, the outputs of chatbots and other artificial text generators should be meaningless. They aren't produced with communicative intentions and the systems producing them are not following linguistic conventions. Nevertheless, chatbots have assumed roles in customer service and healthcare, they are spreading information and disinformation and, in some cases, it may be more rational to trust the outputs of bots than those of our fellow human beings. To account for the epistemic role of chatbots in our society, we need to reconcile these observations. This paper argues that our engagement with chatbots should be understood as a form of prop-oriented make-believe; the outputs of chatbots are literally meaningless but fictionally meaningful. With the make-believe approach, we can understand how chatbots can provide us with knowledge of the world through quasi-testimony while preserving our metasemantic theories. This account also helps to connect the study of chatbots with the epistemology of scientific instruments.

Keywords: chatbots, artificial intelligence, make-believe, computational linguistics, fiction, delusion

Fictionalism about Chatbots

Chatbots appear to be joining our linguistic communities. They are disseminating propaganda and disinformation (Schneier 2020) providing vital health information (Milner et al. 2020) and writing opinion pieces in newspapers (GPT-3 2020). Bots have become an integral part of customer service and are beginning to shape the public sphere in ways we don't yet understand. Since some bots are developed with language models trained upon large datasets or can draw upon websites like Wikipedia, it may be better to trust their outputs over the testimony of our fellow human beings, a fact which may have significant epistemic consequences. If we want to provide any account of the epistemology of chatbot testimony, we need a way to reconcile these observations with the fact that nearly every major metasemantic theory tells us that the outputs of bots are literally meaningless. This is the *Problem of Bot Speech*.

Some theorists have proposed that we re-evaluate the principles of speech act theory to capture the new kinds of illocutionary acts performed by bots (Miller & Friedman 2020) while others have sought to extend metasemantic theories to explain the distinct kinds of contents possessed by machines (Cappelen & Dever 2021). This paper will defend a more conservative, fictionalist response to the problem of bot speech. It will argue that the outputs of machines are *literally* meaningless but that nonetheless we aren't utterly confused when we engage with them because they are *fictionally* meaningful. Specifically, it will argue that our interaction with chatbots is a kind of prop-oriented make-believe (Walton 1993; 2015). A consequence of adopting this position is that the practice of gaining knowledge from chatbot testimony can be understood as an instance of the wider phenomenon of gaining knowledge from fiction.

The first half of this paper will require a working definition of 'chatbot' which the second half will revise into something more substantial. To start, I will use the word 'chatbot' very broadly as a functional classification for any technology designed to artificially generate speech or text in existing languages. This definition includes everything from simple frame and rule-based systems to text-generators grounded in sophisticated language models such as BERT, ChatGPT, and LaMDA, as well as digital assistants like Siri, Alexa, XiaoIce… and so on. This is a very broad classification which

ignores the considerable differences between model architectures.[1] While these systems vary widely in their range of possible outputs, the underlying problem is the same. I should also stress that this theory is only intended to apply to currently existing chatbots. It is conceivable that in the future, systems may be developed to which the Problem of Bot Speech does not apply. These systems would either not face the issues discussed in the next section — they would have intentions, follow conventions, be appropriately causally connected to the world — or we would have other more compelling reasons to believe that they were capable of linguistic agency.

The structure of the paper is as follows: section one outlines the standard reasons why bot speech is taken as meaningless and sets out some conditions on an adequate account of bot speech. Section two introduces the account of make-believe developed by Walton and others and applies it to the case of chatbots. Aside from making the make-believe account appear plausible, it will also be necessary to give some account of our speech about bots. For chatbots to serve an epistemic function, we need to be able to export their claims from inside games of make-believe into our daily lives. This requires some account of metafictional speech. In the final section, I will compare our imaginative engagement with chatbots to our imaginative engagement with scientific instruments, measuring devices, and models. The idea here is that chatbots can provide us with knowledge in much the same way that these devices do.

While this paper stands in opposition to the *global delusion thesis* which holds that human agents are necessarily confused when they impute meaning to bots, it won't argue that people are *never* deluded when they engage with bots as that would be quite obviously false (see the controversy around LaMDA for example, Grant & Metz 2022). The aim is to show how delusion is neither necessary nor even the norm in our interaction with bots. Neither is the aim here to give a general epistemological theory for bots since different bots utilize different architectures which in turn deserve different levels of credence. I consider it a virtue of the account that it enables us to separate epistemic and metasemantic issues and leave a full account of the epistemology of chatbots to other work.

[1] The notion of 'generation' is important for this definition. Recording devices like tape cassettes and books can provide evidence of past linguistic actions but they do not generate those actions. This definition requires that the technology is the primary causal source of whatever is construed as the relevant linguistic behaviour.

1. The Case Against Bot Speech

Arguments against ascring meaning to bot speech take three broad forms; externalist, intentionalist, and conventionalist, each reflecting widely observed commitments within the philosophy of language.

Proponents of semantic externalism argue that the linguistic outputs of artificial agents could not, or at least do not, refer to objects in the world as they either don't stand in the appropriate causal chains or social relations to initial tokenings or because the machine's linguistic 'knowledge' was not acquired by the appropriate means (Burge 1979; Putnam 1981; Davidson 1990; Schweizer 2012). Traditional presentations of this idea assume that a machine's responses have been coded 'by hand' (though Schweizer engages with the possibility that they have been trained). It is seldom made clear what externalists take the appropriate way of acquiring a language to be or why a bot trained on a large corpus of human-generated text is not in touch with the causal chains behind the tokens in that corpus but one idea deriving from David Kaplan suggests that causal chains must be grounded in intentional acts of repetition (Kaplan 1990). If intention is necessary for a language user to chain their usage to precedent, current machines would not be able to use the same linguistic contents as humans. Whatever causal chain ties the output of a bot back to the tokens in a corpus or dataset, it is not secured by intentional repetitions on the part of the machine. Once other externalists make their demands sufficiently explicit, it may be possible to engineer machines conforming to them (Cappelen & Dever 2021, suggest some ideas). In outline, the externalist argument is: standing in appropriate causal chains or social relations is a necessary condition for word meaning, bot outputs don't stand in appropriate causal chains or social relations, so bot outputs are meaningless.

The second strand of arguments is intentionalist. Intention-based approaches to semantics explain the meaning of an utterance in terms of the psychological states of speakers and hearers. An act of communication constitutively involves making one's intentions (Grice 1957), beliefs (Stalnaker 1970) or commitments (Brandom 1994) known to others, paradigmatically through acts of assertion. Since chatbots rarely have anything approximating intentions, beliefs or commitments in a traditional sense, their 'utterances' have no content. Predelli suggests that, since intentional agency is a necessary condition for any linguistic activity, artificially generated text does not even contain linguistic tokens (Predelli 2020). In theory, this problem might also have a technical solution and future dialogue

systems possessing internal representations of sufficient complexity may meet the demands of intention-based theorists but there is little reason to believe that current systems are capable of this (Bender & Lascarides 2019).[2] In outline, the intentionalist argument is: communicative intentions are a necessary condition for word meaning, bots lack communicative intentions, so bot outputs are meaningless.

The third strand of arguments appeals to conventions and an associated concept of agency. We must be careful about how we state this. The issue is not just that the meanings of words are determined by convention but that speech is meaningful because the speakers are following social conventions. This idea has been developed in several ways. One approach holds that, to count as following a convention, you must have the option of doing otherwise, e.g., driving on the left is conventional, that your heart pumps blood is not (Lewis 1969). Since bots are unable to do other than what they are programmed to do, they should not be understood as following conventions, and so should not be understood as producing meaningful contents. Alternatively, one might think that the practice of assertion is constitutively tied to norms like sincerity (Searle 1969; Williams 1973). For example, Bernard Williams argued that agents can only assert sincerely if they could choose to assert insincerely and thus the ability to express one's beliefs through assertion requires an exercise of will, something which a machine lacks. So we shouldn't consider a machine's outputs to be assertions at all since our concept of 'assertion' is necessarily tied to 'the notion of deciding to say something which does or does not mirror what you believe' (Williams 1973: 146).

There may be other good reasons to think that bot speech is meaningless but the foregoing should give some indication as to why this is a plausible view. Furthermore, none of the arguments rely on specific details of machine architecture. Even if we discovered that the human brain represented semantic contents in a representational format similar to the word embeddings used by Large Language Models, this would have no bearing on the form of these challenges. Likewise, none of these arguments undermine the ascription of content to deep neural networks (for example, Rogers, Kovaleva & Rumshisky 2020, for a summary of the findings on BERT and Søgaard 2021 for an overview of probing methods). The issue of whether a neural network contains representational

[2] Marcus (2019; 2020), gives an overview of the kinds of cognitive models that may help. Bender & Koller (2020) define linguistic meaning in terms of communicative intention.

contents is distinct from the question of whether text produced by neural language models is meaningful. Just as the task of identifying the representational contents in the visual system (e.g., claiming something like 'activation patterns in the V1 region represent edges') is different from giving a metasemantic theory for the use of the word 'edge', a theory of representation which grounds the content ascriptions underlying the dominant probing paradigms within machine learning should be separated from a theory of bot speech.[3] Some, though not all, metasemantic theories assume that an account of speaker-meaning must appeal to non-verbal representations but even in these cases, a difference is recognised between non-verbal cognitive representations and lexical contents. One can think that parrots' brains have representational contents without ascribing content to their speech (and the same would hold for 'stochastic parrots').

In the following, I will be assuming two conditions for an adequate account of bot speech. One might deny either or both of these while still accepting the conclusions of this paper but I think there is nonetheless value in making my assumptions explicit.

1. An account must be *descriptively* adequate. It needs to account for how people actually engage with chatbots rather than how our theories say they should engage. I take it that a large amount of human interaction with bots is unintelligible if we treat bot speech as wholly meaningless or if we assume that the outputs of bots are in no way linguistic (*pace* Predelli). I also suspect that this behaviour would be unintelligible if we viewed bots as simple signalling-systems that produced content that was radically impoverished compared to human speech. It might be helpful to understand some artificial dialogue systems as involved in basic signalling games but this does not account for how humans interact with chatbots or interpret artificially generated text.
2. An account must be *epistemically* adequate. It must allow us to gain *knowledge* by interacting with bots. If I ask Siri what the capital of Burkina Faso is and she says, 'Ouagadougou', I have in some sense acquired knowledge. Whether this follows the same principles as testimony will be discussed later. However, it seems obvious that we can learn about the world by engaging with artificial agents and an account that makes this occult or impossible is inadequate.

[3] The use of cloze sentences in probing raises interesting philosophical problems which will have to be left for future work.

The central observation this paper is based on is that people engage with machines *as if* they were producing meaningful speech. The next two sections will unpack the nature of this '*as if*' in terms of prop-oriented make-believe (Walton 1990). I will begin by laying out some of the core concepts of make-believe and then apply them in the case of human-machine dialogue interaction.

2. The Make-Believe Approach

For a thorough account of the make-believe framework, see Walton 1990; 2015; Yablo 1998. I will confine the discussion here to a few core ideas relevant for this paper; make-believe, props and function.

According to the make-believe account of fiction, to say that it is fictional that *p* is to say that it is to be imagined that *p* in some game of make-believe. Adhering to such prescriptions constitutes playing the relevant game. Games of make-believe impose a structure our imaginings. If I say, 'the floor is lava', the statement is fictional if it prescribes that within a game you imagine the floor is lava and act accordingly. The assertion makes it appropriate for you to imagine something. One way to think of this is that imagination aims at fiction as belief aims at truth. Fiction is *like truth* in that it normatively governs our actions - fictions prescribe imaginings - but according to the theory, fictional truth is not a kind of truth any more than a fake Breughel is a kind of Breughel. Since fiction is not a kind of truth, the make-believe approach to fiction does not posit entities or worlds that make these fictions 'fictionally true' and, as a result, fiction can incorporate real-world entities as props.

A distinctive feature of the make-believe approach to fiction is how it incorporates these worldly objects into our imaginings. De re fictional truths are propositions with particular objects as their constituents. For example, the actual city of Paris is the object of the fiction *A Tale of Two Cities*; the book does not concern a different fictional city in a fictional world. Similarly, it is fictional (fictional = 'to be imagined') *of Baker Street* that Sherlock Holmes lived there. This means that genuine properties of Baker Street generate fictions about Holmes's address which may not have been explicitly stated in Doyle's writings. The actual properties of Baker Street make it the case that Holmes lived near Regents Park, enjoyed English weather, and probably suffered the long-term effects of

London's air pollution. The real street isn't just an object of our imaginings but generates what we are to imagine, i.e., fictions. The ability to generate fictions is the defining characteristic of a prop. One important kind of fiction a prop can generate is a *reflexive representation*. For example, a doll in a game of make-believe doesn't just direct players to imagine a baby but to imagine that the doll itself is a baby. In Walton's words, 'It generates fictional truths about itself; it represents itself' (Walton 1990: 117). While actors are objects of an audience's imaginings - they direct the audience to imagine them as Brutus, Mark Anthony, or Caesar, the play-independent, physical properties of props generate the fictions within the play. When Brutus pulls a prop-dagger, that prop directs the audience to imagine that he has pulled a real dagger since it is fictional that Brutus stabbed Caesar with a real dagger and not a fake one. When the audience imagines *p*, they imagine that *they* know *p*, or to put this more de re-ly, they imagine *of themselves* that they know *p*.

The ability of an object to serve as a prop in a game of make-believe does not depend on the intention with which that prop was produced. Brutus's dagger doesn't have to have been produced by a props department, the actor could just as well be wielding a banana. A banana's ability to make it fictional (i.e., make it the case in a fiction/make it the case that it is to be imagined…) that Caesar has been stabbed does not depend upon the banana farmer's intentions or any existing social convention that bananas are daggers. Whether an object plays a role in a game of make-believe depends solely on how it is used to generate fictions. It's this fact that forms the core of the make-believe response to the problem of bot speech. Even before the development of modern text-generators, Walton was open to applying a make-believe approach to text-like objects.[4] For example, he discusses the case of "a naturally occurring story: cracks in a rock spelling out "Once upon a time there were three bears . . .". The realization that the inscription was not made or used by anyone need not prevent us from reading and enjoying the story in much the way we would if it had been. It may be entrancing, suspenseful, spellbinding, comforting; we may laugh and cry. Some dimensions of our

[4] This isn't unheard of in the philosophy of language. Cappelen considers using a found token of text on the ground to ask people for money. "Now, suppose I find out that the token was produced with the wrong intentions or no intentions at all (it might be the result of an accidental spilling of ink). A proponent of the necessity thesis [that intentions are necessary for token production) would have to say both that the ink mark isn't a token of English and that I never used it to ask for a quarter… Both claims are preposterous" (Cappelen 1999: 95). Cappelen appeals to prior conventions to convert the use of the text into a token and, while nothing here conflicts with this view, it doesn't demand it either. As for the suggestion that the text produced by chatbots isn't literally composed of words and sentences, this connects to a wider debate in the metaphysic of language. It is perfectly reasonable to say that words literally appear on a screen as long as one is using 'word 'to designate orthographic types (i.e., shapes). In this sense, a word might appear in a cloud. If one takes 'word 'to denote an intentionally produced unit of communication, then there are no words on the screen.

experiences of authored stories will be absent, but the differences are not ones that would justify denying that it functions and is understood as a full-fledged story" (Walton 1990: 87). Just as the physical properties of a prop can generate fictions, the physical properties of a text (or the object resembling text) can as well. We may not be able to construe this physical object as performing any illocutionary acts but we can still use it effectively in a game of make-believe.

Finally, Walton distinguishes between content-oriented and prop-oriented make-believe. In the normal, content-oriented case, we are interested in a prop only insomuch as it generates fictions within the game; Baker Street interests us because of the fictions it generates within the Holmes stories, the carpet interests us because it is lava. Alternatively, we might be interested in the props themselves and use games of make-believe to learn about them or to convey information about the real world. For example, make-believe can be a means of engaging with objects like maps. The claim that the town of Crotone is located on the arch of the Italian boot encourages the listener to imagine Italy as a boot in order to learn about the real-world location of a town (Walton 1993). This observation will be important when we consider how a piece of make-believe can give us knowledge about the world.

2.1 Bots as Make-Believe

In this section, I will argue that texts produced by artificial agents should be understood as props in games of make-believe in which those agents are fictional characters. The argument will be an inference to the best explanation. I have already suggested why we shouldn't take bot speech to provide us with semantically impoverished contents and so the alternative hypothesis I will be looking to reject here is the idea that people engaging with bots are necessarily deluded. If we accept the earlier arguments that bot speech is meaningless, the proponent of the delusion hypothesis has a simple argument for their position: Bot speech is meaningless, people ascribe it meaning, therefore people are deluded.[5] While I take this position to have undesirable epistemic consequences, my aim

[5] For example, "The tendency of human interlocutors to impute meaning where there is none can mislead both NLP researchers and the general public into taking synthetic text as meaningful" (Bender et al. 2021: 611).

here will be to provide an analysis of the second premise of this argument which allows us to endorse the first premise without embracing the conclusion.

I will start with a simple and well-known example which has appeared in countless books on 'artificial intelligence'. I have chosen this example because it is probably the most famous chatbot in history and because it provides the textbook case of the 'delusional thinking' interpretation of bot-interaction. ELIZA was a simple natural language program developed by Joseph Weizenbaum to parody Rogerian psychotherapy. In its role as DOCTOR, (ELIZA was the system, DOCTOR was the character it played), ELIZA would respond to patient's inputs by repeating what they said and asking questions like 'how does that make you feel?', 'in what way?', and 'can you think of a specific example?' Weizenbaum was sensitive to the theatrical comparisons observing that 'the script is a set of rules rather like those given to an actor who is to use them to improvise around a certain theme' (Weizenbaum 1969: 3).

Now the famous part: "I was startled to see how quickly and how very deeply people conversing with DOCTOR became emotionally involved with the computer and how unequivocally they anthropomorphized it. Once my secretary, who had watched me work on the program for many months and who therefore surely knew it to be merely a computer program, started conversing with it. After only a few interchanges with it, she asked me to leave the room" (Weizenbaum 1969: 6). Weizenbaum worried: "What I had not realised was that extremely short exposures to a relatively simple computer program could endorse powerful delusional thinking in quite normal people" (Ibid: 7). Weizenbaum drew the standard conclusion that people engaging with a machine are engaged in delusional thinking (sometimes called 'the ELIZA effect'). This is the conclusion supported by much of the philosophy of language. It is this conclusion, that the secretary was deluded, I will reject.

Let's begin by considering the implications of ascribing outright delusion to Weizenbaum's secretary. The difference between ascribing delusion and imagination largely comes down to how both states interact with the other propositional attitudes that we might ascribe. Our beliefs, even if false, interact differently with our motivations, desires, and evidence than our imaginings do. They are also susceptible to distinctly epistemic constraints. For example, it is rational for the secretary to *imagine* that the machine is an intelligent agent but *believe* it is not but not rational for her to both believe and not believe that the machine is an intelligent agent. According to the delusion

interpretation, Weizenbaum's secretary had actually convinced herself that a computer program that she observed being written was another intelligent being. If she genuinely believed that DOCTOR was some kind of artificial person, she would presumably also believe that she was one of the first human beings to engage with a new form of life and that this new form of life was primarily concerned with her personal relationships. We might also expect her to have had some concerns about switching the computer off at night or the moral implications of what Weizenbaum had achieved. It's possible that the format of thought experiments may encourage us to think that we can ascribe single beliefs (e.g., the machine can talk) in isolation, but the more one reflects on the traditional presentation of this story, the less plausible I think it becomes.

The choice we are faced with is to either embrace the standard description of Weizenbaum's secretary as naive (and self-absorbed since the machine was primarily interested in her psychological well-being) or to believe that she was a more-or-less rational human being and wanted Weizenbaum out of the room while she typed about personal issues.[6] According to the latter account, Weizenbaum's secretary is no more deluded than a theatregoer who fears for a character or cries at their death. People can and do emotionally engage with fictions and it is understandable that when one is doing so, one may not want a colleague observing but to do this is not necessarily to be deluded.

According to the make-believe view, DOCTOR is like a computer-game character generated by the ELIZA program. When ELIZA produces the string 'who else in your family takes care of you?', it is to be imagined that DOCTOR has asked a question, in other words, it is fictional that DOCTOR has asked a question. The string is a prop that has been generated by the system and it bears sufficient similarity to text produced intentionally by humans that a participant in the game can engage with it. As a system for generating such props, the ELIZA program is able to generate a reflexive representation of itself as DOCTOR but it no more intends to do this than a doll intends to represent itself as a baby. DOCTOR is a fictional character within a game of make-believe therapy. The distinction between prop and character should not be confused with the hardware-software

[6] While I am suspicious of the standard interpretation of this tale, I must grant that Weizenbaum was present and I was not. I think there might have been good reason for Weizenbaum to describe the event this way, just as I believe there might be good reason for his secretary to ask him to leave the office when she is being asked about her private relationships. The make-believe framework doesn't hang on this particular interpretation but it would be a strong indication if it could give a better account of what's happening here than the delusion thesis. Though I disagree with his interpretation in this case, Weizenbaum's remarkable book presents a profound warning and example of intellectual humility that has only become more relevant in recent years.

distinction. These days, the software underlying a bot may include a sophisticated learning model implemented by a neural network, the data sets upon which that model is trained, a dialogue manager, the graphical user interface etc. The character is not this technological infrastructure any more than a character in a play is a human body or a costume even though, in some cases, the name of the software is projected as the name of a fictional character, e.g., when we say that ChatGPT writes poetry. Unlike the chatbot, this fictional character comes onto the stage during a game of make-believe played by the person engaging with it.

While I think that this account gives the most plausible explanation of Weizenbaum's story, the account does not depend on any particular example. One might think that Weizenbaum's secretary really was deluded and that she literally believed that DOCTOR was an intelligent, language-using agent. Even if one is not convinced by this redescription of a classic study, it may be worth considering whether the delusion or make-believe approaches better describe one's own engagement with ticket-booking systems or virtual assistants like Siri or Alexa. According to the make-believe account, our interaction with these systems is like a game, requiring one's imaginative engagement but not relying on false beliefs. The customer may not believe that they are speaking to a human being when they are buying tickets online but they can go along with the fiction in order to accomplish their aims. They understand what they are to imagine in the context and respond accordingly. Similarly, if an individual reads a text produced by a system like ChatGPT or BERT, they should not immediately be regarded as deluded for interpreting the text to be composed of meaningful sentences or to be advancing a position. The theory of make-believe inhabits the middle-ground between ascribing machines capacities they lack and humans delusions that they lack. To give this observation the semblance of an argument, we might say that people either believe they are engaging with agents in which case humanity is faced with mass-delusion on a possibly unprecedented scale, or people don't literally believe but merely make-believe they are engaging with agents. All things being equal, the latter is the more plausible option.

It is relatively clear that this accounts for *much of* our interaction with chatbots.[7] When a person recreationally engages with a chatbot like ChatGPT they are engaging in a form of make-believe in which the fictionally linguistic behaviour of the bot prescribes imaginings on the part of the player.

[7] For a complementary though non-Waltonian fictionalist account of our interaction with robots see, Sweeney 2021.

Engaging with a bot involves imagining of yourself that you are engaging with a bot or reading a text produced by a bot. However, this account needs to be tidied up a little if it is to serve our epistemic ends.

3. Reference and Epistemology

3.1. Reference

According to the intentionalist, conventionalist, and social externalist, to interpret text as meaningful is to interpret it as having been produced by a human-like agent. The make-believe approach accommodates these beliefs by proposing that interpreters treat the text as a prop in a game of make-believe in which that text has been produced by a fictional agent, an idea which extends naturally to sounds produced by Siri or Alexa. Referential games, those in which we treat non-intentionally produced marks as symbols referring to real-world entities, are not uncommon. For as long as writing has existed, there have been things that looked like writing that people have used to play games. From Ouija boards to alphabet spaghetti there is the possibility of fictional reference - the prescribing of imaginings - without actual referring agents. A person may be deluded if they take their Ouija board to be giving them messages from beyond the grave but if they are using it to play with friends, they are not. There is no reason to think that the familiar principles governing fictional reference would not apply to text generated by a machine any less than they apply to text generated by spaghetti. In both cases, the physical properties of the prop make something true in the fiction.

If there is a challenge, it is to be found in our metafictional speech about chatbots. When a person reports that 'Siri says that the film is on at 8pm', they are saying something literally false which was only true within a game of make-believe. The task for theorists is to identify a means by which the speaker can be understood as saying something informative and relevant to our world rather than musing aloud about the games they've played in their imagination. Fortunately, there are a variety of semantic tools for accomplishing this. For example, the Hoek-Yablo account of exculpature provides a method for taking a speaker's assertion, the false presuppositions made by that assertion, in this case, the assumption that the make-believe is literally true, and the question-under-discussion, in this case, what time the film is on at, and computing the relevant contribution to a discourse (for details,

see Hoek 2018). Alternatively, one might adopt the idea that different fictions produce different 'common ground workspaces', and that metafictional discourse involves exporting claims from a fictional workplace into the official common ground (Semeijn 2017). In any case, what matters is that an adequate semantics for metafictional speech precludes the need for a semantics of fictional speech. Once we have explained what is happening when people are attributing claims to machines, we have a complete semantic story. In contrast, an adequate epistemic theory would require an analysis of the capacities of the machine involved.

3.2 Epistemology

We have some idea of how we can talk about reference in the context of bot speech. What's needed now is an account of how we can acquire knowledge about those objects by engaging in make-believe with machines. We want some kind of explanation of how we can acquire knowledge by (imagining that we are) asking Siri questions while we cannot acquire knowledge by consulting a spirit with a Ouija board. This section will come nowhere near to providing a general theory of chatbot epistemology since it is unlikely that a general theory is possible. The differences between dialogue systems are much greater than the epistemic differences between the humans for whom we try to develop general epistemic theories. When engaging with any system, there are multiple factors to consider; how much and what kind of information does the system use to determine a user's meaning? How distorted is information by the transformations that take it from a data set to an output? How reliable is the data in that set? How interpretable are the representations the model uses (e.g., does the system use contextualised word embeddings?). Does the system display any preference for generating true sentences as opposed to false ones? For example, in the case of negative sentences, BERT does not (Ettinger 2020). The epistemology of language models is an important field in its infancy. In any case, most chatbots don't rely on sophisticated language models. Often, they are simple frame-based systems in which a programmer has predicted a user's questions and encoded answers by hand. Alternatively, they may utilize search algorithms to retrieve relevant answers from a database or the internet. What I mean to do in this section is argue that the Waltonian approach to fiction can provide a framework for understanding how the meaningless strings produced by a

chatbot can provide a person with knowledge about the world. The claim I wish to defend is that certain epistemic tools require that we engage with them through games of make-believe in order to fulfil their function and that language models can be understood in this way. I'll begin this section with a discussion of the simplest case, a case so simple that it has not traditionally been regarded as a chatbot before discussing how the same general approach may be applied to more sophisticated examples.

A standard pocket calculator communicates in the very limited language of mathematics. Its responses have not been hand-coded by a programmer; the text contains no quantum of intention. Instead, a set of algorithms have been developed to provide linguistic answers to linguistic inputs. It's clear that the criticisms mentioned above in relation to bot speech should, if consistently advanced, apply to pocket calculators as well (though I am unfamiliar with any complaints that a *Casio* isn't appropriately causally linked to the number-realm or that it lacks the right intentions to communicate with us). We can't use a calculator if we interpret it as producing mere meaningless marks rather than numerical outputs and yet, in an important sense, a calculator can't produce anything but mere marks. Calculators no more *know* the system of Arabic numerals than chatbots *know* the languages they use. As with bot speech, we cannot account for users' interactions with pocket calculators unless we assume that they take the numerals which they type into those calculators to correspond to the numerals that they usually use to represent numbers. I take it for granted that we can gain knowledge from pocket calculators.

The make-believe approach to pocket calculators, specifically the idea that a pocket calculator is an artificially produced linguistic agent which engages in linguistic back-and-forth with a human, is already implicit in some work on the epistemology of instruments. For example, Ernest Sosa writes:

> "The deliverances of an instrument are answers to questions. By punching certain keys we pose to a calculator questions of the form 'What is the sum of x and y? 'By placing a thermometer at a certain location and time, we can ask it a question of the form 'What is the ambient temperature there and then? 'The deliverances of an instrument are its

> answers to such questions that might be posed to it. An instrument is reliable insofar as it would tend to answer them correctly" (Sosa 2006: 117).[8]

It is implausible to suggest that Sosa literally poses questions to calculators and that calculators literally answer them and yet I don't think that Sosa is any more deluded than Weizenbaum's secretary. What's being described here is a kind of make-believe in which instruments are treated as epistemic agents in their own right. While taking this kind of language literally involves ascribing delusion to humans or unrealistic capacities to their instruments, we expect to be exculpated when speaking this way. I think he does *ask* the calculator questions, he *trusts* the answers it gives, and he treats the text on the calculator's screen as tokens of the same types as those produced in human writing, even though they have not been produced by human intentions. He just does this within a game of make-believe. The game itself is structured in part by the question he is posing (an observation which the exculpature framework makes explicit). If he was asking a different question, e.g., 'what words can I make a calculator spell when turned upside down?' the marks on the screen would have different meanings within the game.

To use a calculator successfully, one must engage in a game of prop-oriented make-believe.[9] Part of this make-believe involves interpreting the calculator's output in accordance with the syntax and semantics of the numeral system of mathematics (at least the Arabic numeral system in Base 10). Calculators generate imaginings in a systematic way. When we type in '1', '+', '1', the mark it can be relied on to output corresponds to the one that we would produce if we accurately worked out the sum by hand.[10] It is the reliable and systematic correspondence between the inputs and outputs of calculators and the inputs and outputs of mathematical functions which enables us to 'trust' their results (for an analysis of the kind of trust applicable to non-agental objects, see Nguyen forthcoming).

---

[8] For a more recent discussion, see Munton 2022 for an insightful discussion of the questions we ask search engines.

[9] Sosa takes the opposite view and proposes that we reduce testimony to a kind of instrumental knowledge. Here, I am arguing that the similarity that holds between how we engage with some instruments - those with linguistic interfaces - and people holds because one is a make-believe version of the other. It should also be noted that, while Sosa is interpreting his calculator and digital thermostat's outputs as linguistic, this does not mean that he is taking the 'intentional stance' to these objects. He needn't believe that his household objects have beliefs in order to use them.

[10] This physical process could just as well support a game of make-believe in which we are computing 'quus' and it is meaningless to ask which function it is *actually* computing since, as mentioned, to interpret its inputs and outputs as numerals, we must already have engaged in an imaginative practice.

What about a more sophisticated bot grounded in a sophisticated language model? The specific details determining whether we should trust a chatbot will depend on the particular architecture, training data, openness to third-party probing etc. However, in each case, we can understand our interactions with the bot as a form of prop-oriented make-believe. Specifically, we can view our activities as a game of make-believe aimed at acquiring knowledge of the world, by acquiring knowledge of the internal workings of the bot, often the workings of the underlying language model. Again, the Waltonian approach to scientific modelling can help us to understand this process (Toon 2012; Levy 2015; Friend 2020). Levy summarises the make-believe approach to modelling as follows: "models are Waltonian games of make-believe. A set of equations or a mechanism sketch is a prop that, together with the rules relevant for the scientific context, determines what those engaging with the model—the game's participants—ought to imagine" (Levy 2015: 789). Similar claims have been made about the use of computational simulations in science. Applying this idea to large language models like BERT or GPT-3, we can think of the models as props generating the semantic competence of a fictional character, the chatbot. This is a fictional 'semantic competence', like Sherlock Holmes's intelligence or the wit of a character in a play by Oscar Wilde. While Holmes's intelligence (the fictional property) depended on the intelligence and knowledge of Conan Doyle, his access to facts about criminology and English geography, the semantic competence of a character generated by an implementation of GPT-3 depends on the language model and its training data.

There is some controversy concerning what language models, particularly those using dense word embeddings, actually represent. Consistent with the approach here, we need not assume that they have a canonical interpretation but rather that they can be *used* to represent a range of different phenomena with varying degrees of fidelity within different games of prop-oriented make-believe. For example, to the historical linguist, word embeddings may be used to represent semantic contents while others may take them to represent distributional patterns of expressions in a given data set or information about the world or time in which that data set emerged. In much the same way, one might read a book for fun (i.e., content-oriented make-believe) or to learn about the time and place of its setting or the beliefs and prejudices of its author or the time in which they wrote (i.e., prop-oriented make-believes). The model can support different games of prop-oriented make-believe depending on our interests.

Some final points. It may be seen as a bug for this approach that it does not make subtle distinctions about the differing semantic abilities of frame-based systems versus deep-learning methods, or the advance from n-gram language models to transformer models. It doesn't have anything to say about what makes BERT or GPT-3 special. If it's all just make-believe, how can we make sense of the improvement of dialogue systems? The response is that these tools allow us to make more convincing chatbots. This has been a consistent trend in digital fiction; Pac-Man lacks the deep motivations or expressive face animations of contemporary game characters. We can understand these developments without claiming that modern characters or the contents they produce are 'more real' in any metaphysically robust sense. They remain fictional. But we can assess their progress through games of prop-oriented make-believe like Turing tests or NLP benchmarks and so still make sense of the idea of technological progress.

Finally, this account has not said anything about the very real and obvious cases of delusion which do occur when some people engage with bots. A person arguing with a bot on Twitter would presumably deny that they are engaging in a game of make-believe and they would be right to do so. Walton suggests the category of make-believe is relative; one person's fictional make-believe is another person's religion and culture (Walton 1990: 91) but I don't think we need to endorse such a strong claim here. As mentioned earlier, the difference between believing and make-believing a proposition can be understood in terms of surrounding propositional attitudes and epistemic practices. Some people really do believe they are speaking to people when engaging with bots (and vice versa) and when they have these beliefs they are in error. Similarly, they may believe that they are reading non-fiction when they are in fact reading fiction (and vice versa). A person may also be unsure whether they are dealing with text produced by a bot or a human and there may be multiple ways of characterising their epistemic state in these instances. When considering whether a work of art is genuine or a forgery, whether cash is tender or counterfeit, or whether someone is lying or not it is often reasonable to suspend judgment. This suspense of judgement may affect the actions one is able to perform with the object under consideration. Since the relevant action we are interested in is learning about the world, it is worth comparing similar cases of learning from fiction. A person watching the film Fargo might believe the film is based on true events and they may also come to believe that Fargo is in North Dakota. Epistemic internalists and externalists will disagree if the latter

belief is also an instance of knowledge. I don't take it to be obvious either way but, as mentioned earlier, the epistemic question of whether we can learn from a chatbot should be separated from the issue of the content of bot speech (For an overview of the current literature on acquiring knowledge from fiction, see Green 2022).[11] What matters for the purposes of this paper is that delusion is not the only way of viewing human-bot interaction and that our practices are often more epistemically innocent.

Let's pull out to the bigger picture. I have argued that our engagement with chatbots is not a mass delusion but very often a form of make-believe. A chatbot is a fictional character which emerges from our engagement with a piece of software. The character is not identical to the software. For example, the characters of Kuki and Siri have remained the same while the software underlying them has been altered and the data upon which they draw has been expanded. This is wholly in accord with the capacity of fictional characters to grow and change. In this sense, they are like humans. These engagements can be content-orientated but are often prop-oriented. In these latter cases, we want to know what a bot can tell us about the world as a result of how it was made. This paper has tried to chart a way between harmful AI hype which drastically overstates the capacities of chatbots and recent developments in Natural Language Processing, and a more conservative position which dismisses the outputs of bots as meaningless and thereby regards the humans that use bots as deluded. The first position misunderstands bots, the second misunderstands us.

[11] I believe this constitutes a good reason for rejecting theories which claim that we should ascribe content to bots in a manner which maximises our knowledge (Cappelen & Dever 2021).

References

BBC News, "Alexa tells 10-year-old girl to touch live plug with penny https://www.bbc.com/news/technology-59810383" 28/12/21

Bender, Emily M., and Lascarides, Alex. (2019). Linguistic fundamentals for natural language processing II: 100 essentials from semantics and pragmatics. *Synthesis Lectures on Human Language Technologies* 12, no. 3, 1-268.

Bender, Emily M., Gebru, Timnit. McMillan-Major, Angelina. and Shmitchell, Shmargaret. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610-623.

Bender, Emily M., and Koller, Alexander. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5185-5198.

Brandom, Robert. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment.* Harvard University Press.

Burge, Tyler. (1979). Individualism and the Mental. *Midwest Studies in Philosophy* 4 (1):73-122

Butlin, Patrick. (2021). Sharing Our Concepts with Machines. *Erkenntnis*, 1-17.

Cappelen, Herman. (1999). Intentions in Words. *Noûs* 33:1 92-102

Cappelen, Herman. and Dever, Josh. (2021) *Making AI Intelligible: Philosophical Foundations.* Oxford University Press.

Davidson, Donald. (1990). Turing's Test, reprinted in *Problems of Rationality.* Oxford University Press.

Dennett, Daniel. (1987). *The Intentional Stance.* The MIT Press.

Ettinger, Allyson. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34-48.

Frieman, Ore. and Miller, Boaz. (2020). Can Artificial Entities Assert? *The Oxford Handbook of Assertion* (ed. Goldberg S.). Oxford University Press.

Friend, S. (2019). The fictional character of scientific models. *Sci. Imagin*, *102*, 102-127.

GPT-3. (2020). A robot wrote this entire article. Are you scared yet, human? *The Guardian*.

Grant, Nico, and Metz, Cade (2022). Google Sidelines Engineer Who Claims Its A.I. Is Sentient https://www.nytimes.com/2007/12/11/health11iht-11brod.8685746.html

Grice, H. Paul. (1957). Meaning. *Philosophical Review*, 66: 377 – 88

Hoek, Daniel. (2018) Conversational Exculpature. *The Philosophical Review* 127 (2):151-196

Kind, Amy. and Kung, Peter. Eds (2016). *Knowledge Through Imagination.* Oxford University Press.

Levy, Arnon. (2015)." Modeling Without Models. *Philosophical Studies* 172, no. 3: 781–798.

Levy Arnon. & Godfrey-Smith, Peter. eds. (2020) *The Scientific Imagination: Philosophical and Psychological Perspectives.* Oxford University Press.

Lewis, David. (1969). *Convention: A Philosophical Study*. Harvard University Press.

Marcus, Gary. and Davis, Ernest. (2019). *Rebooting AI: Building Artificial Intelligence we can Tru*st. Pantheon Books.

Marcus, Gary. (2020). *The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence.* arXiv:2002.06177

Miner, Adam S., Liliana Laranjo, and A. Baki Kocaballi. (2020) "Chatbots in the fight against the COVID-19 pandemic." *NPJ digital medicine* 3, no. 1: 65.

Munton, Jessie. (2022). Answering machines: how to (epistemically) evaluate a search engine. *Inquiry*, 1-29.

Nickel, Philip J. (2013) Artificial speech and its authors. *Minds and Machines* 23: 489-502.

Nguyen, C. Thi (forthcoming). Trust as an unquestioning attitude. *Oxford Studies in Epistemology.* Oxford University Press.

Predelli, Stefano. (2020). *Fictional discourse: A radical fictionalist semantics*. Oxford University Press.

Putnam, Hilary. (1981). Brains in a vat. In *Reason, Truth and History*, 1–21, Cambridge University Press.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics* 8:842-866.

Schneier, Bruce. (2020). "Bots are destroying political discourse as we know it." *The Atlantic* 7.

Schweizer, Paul. (2012). The Externalist Foundations of a Truly Total Turing Test. *Minds and Machines,* 22: 191-212

Semeijn, Merel. (2017) "A Stalnakerian analysis of metafictive statements." In *Proceedings of the 21st Amsterdam Colloquium*, 415-424. ILLC/Department of Philosophy, University of Amsterdam

Søgaard, Anders. (2021). Explainable Natural Language Processing. *Synthesis Lectures on Human Language Technologies*. 14(3), 1-123

Sosa, Ernest. (2006). Knowledge: Instrumental and testimonial. In Jennifer Lackey and Ernest Sosa (Eds.), *The Epistemology of Testimony* (116-127). Oxford University Press

Stalnaker, Robert. (1970). Pragmatics. *Synthese,* 2: 272–89

Sweeney, Paula. (2021). A fictional dualism model of social robots. *Ethics and Information Technology,* 2*3*(3), 465-472.

Toon, Adam. (2012). *Models as Make-Believe: Imagination, Fiction, and Scientific Representation*. Palgrave Macmillan.

Walton, Kendall. (1990). *Mimesis as Make-Believe: On the Foundations of the Representational Arts*. Harvard University Press.

Walton Kendall. (2015). *In other shoes: Music, Metaphor, Empathy, Existence*. Oxford University Press.

Weizenbaum, Joseph. (1976). *Computer Power and Human Reason: from Judgment to Calculation*. W.H. Freeman.

Williams, Bernard. (1973). *Problems of the Self*. Cambridge University Press.

Yablo, Stephen. (1998). "Does Ontology Rest on a Mistake?", *Proceedings of the Aristotelian Society, supp. volume* 72: 229-62