Ethics and Information Technology (2024) 26:38
https://doi.org/10.1007/s10676-024-09775-5

ORIGINAL PAPER

ChatGPT is bullshit

Michael Townsen Hicks[1] · **James Humphries**[1] · **Joe Slater**[1]

Abstract
Recently, there has been considerable interest in large language models: machine learning systems which produce human-like text and dialogue. Applications of these systems have been plagued by persistent inaccuracies in their output; these are often called "AI hallucinations". We argue that these falsehoods, and the overall activity of large language models, is better understood as *bullshit* in the sense explored by Frankfurt (On Bullshit, Princeton, 2005): the models are in an important way indifferent to the truth of their outputs. We distinguish two ways in which the models can be said to be bullshitters, and argue that they clearly meet at least one of these definitions. We further argue that describing AI misrepresentations as bullshit is both a more useful and more accurate way of predicting and discussing the behaviour of these systems.

Keywords Artificial intelligence · Large language models · LLMs · ChatGPT · Bullshit · Frankfurt · Assertion · Content

Introduction

Large language models (LLMs), programs which use reams of available text and probability calculations in order to create seemingly-human-produced writing, have become increasingly sophisticated and convincing over the last several years, to the point where some commentators suggest that we may now be approaching the creation of artificial general intelligence (see e.g. Knight, 2023 and Sarkar, 2023). Alongside worries about the rise of Skynet and the use of LLMs such as ChatGPT to replace work that could and should be done by humans, one line of inquiry concerns what exactly these programs are up to: in particular, there is a question about the nature and meaning of the text produced, and of its connection to truth. In this paper, we argue against the view that when ChatGPT and the like produce false claims they are lying or even hallucinating, and in favour of the position that the activity they are engaged in is bullshitting, in the Frankfurtian sense (Frankfurt, 2002, 2005). Because these programs cannot themselves be concerned with truth, and because they are designed to produce text that *looks* truth-apt without any actual concern for truth, it seems appropriate to call their outputs bullshit.

We think that this is worth paying attention to. Descriptions of new technology, including metaphorical ones, guide policymakers' and the public's understanding of new technology; they also inform applications of the new technology. They tell us what the technology is for and what it can be expected to do. Currently, false statements by ChatGPT and other large language models are described as "hallucinations", which give policymakers and the public the idea that these systems are misrepresenting the world, and describing what they "see". We argue that this is an inapt metaphor which will misinform the public, policymakers, and other interested parties.

The structure of the paper is as follows: in the first section, we outline how ChatGPT and similar LLMs operate. Next, we consider the view that when they make factual errors, they are lying or hallucinating: that is, deliberately uttering falsehoods, or blamelessly uttering them on the basis of misleading input information. We argue that neither of these ways of thinking are accurate, insofar as both lying and hallucinating require some concern with the truth of their statements, whereas LLMs are simply not designed to accurately represent the way the world is, but rather to

✉ Michael Townsen Hicks
Michael.hicks@glasgow.ac.uk

James Humphries
James.Humphries@glasgow.ac.uk

Joe Slater
Joe.Slater@glasgow.ac.uk

[1] University of Glasgow, Glasgow, Scotland

Published online: 08 June 2024

give the impression that this is what they're doing. This, we suggest, is very close to at least one way that Frankfurt talks about bullshit. We draw a distinction between two sorts of bullshit, which we call 'hard' and 'soft' bullshit, where the former requires an active attempt to deceive the reader or listener as to the nature of the enterprise, and the latter only requires a lack of concern for truth. We argue that at minimum, the outputs of LLMs like ChatGPT are soft bullshit: bullshit–that is, speech or text produced without concern for its truth–that is produced without any intent to mislead the audience about the utterer's attitude towards truth. We also suggest, more controversially, that ChatGPT may indeed produce hard bullshit: if we view it as having intentions (for example, in virtue of how it is designed), then the fact that it is designed to give the impression of concern for truth qualifies it as attempting to mislead the audience about its aims, goals, or agenda. So, with the caveat that the particular kind of bullshit ChatGPT outputs is dependent on particular views of mind or meaning, we conclude that it is appropriate to talk about ChatGPT-generated text as bullshit, and flag up why it matters that – rather than thinking of its untrue claims as lies or hallucinations – we call bullshit on ChatGPT.

What is ChatGPT?

Large language models are becoming increasingly good at carrying on convincing conversations. The most prominent large language model is OpenAI's ChatGPT, so it's the one we will focus on; however, what we say carries over to other neural network-based AI chatbots, including Google's Bard chatbot, AnthropicAI's Claude (claude.ai), and Meta's LLaMa. Despite being merely complicated bits of software, these models are surprisingly human-like when discussing a wide variety of topics. Test it yourself: anyone can go to the OpenAI web interface and ask for a ream of text; typically, it produces text which is indistinguishable from that of your average English speaker or writer. The variety, length, and similarity to human-generated text that GPT-4 is capable of has convinced many commentators to think that this chatbot has finally cracked it: that this is real (as opposed to merely nominal) artificial intelligence, one step closer to a human-like mind housed in a silicon brain.

However, large language models, and other AI models like ChatGPT, are doing considerably less than what human brains do, and it is not clear whether they do what they do in the same way we do. The most obvious difference between an LLM and a human mind involves the *goals* of the system. Humans have a variety of goals and behaviours, most of which are extra-linguistic: we have basic physical desires, for things like food and sustenance; we have social goals and relationships; we have projects; and we create physical objects. Large language models simply aim to replicate human speech or writing. This means that their primary goal, insofar as they have one, is to produce human-like text. They do so by estimating the likelihood that a particular word will appear next, given the text that has come before.

The machine does this by constructing a massive statistical model, one which is based on large amounts of text, mostly taken from the internet. This is done with relatively little input from human researchers or the designers of the system; rather, the model is designed by constructing a large number of nodes, which act as probability functions for a word to appear in a text given its context and the text that has come before it. Rather than putting in these probability functions by hand, researchers feed the system large amounts of text and train it by having it make next-word predictions about this training data. They then give it positive or negative feedback depending on whether it predicts correctly. Given enough text, the machine can construct a statistical model giving the likelihood of the next word in a block of text all by itself.

This model associates with each word a vector which locates it in a high-dimensional abstract space, near other words that occur in similar contexts and far from those which don't. When producing text, it looks at the previous string of words and constructs a different vector, locating the word's surroundings – its context – near those that occur in the context of similar words. We can think of these heuristically as representing the meaning of the word and the content of its context. But because these spaces are constructed using machine learning by repeated statistical analysis of large amounts of text, we can't know what sorts of similarity are represented by the dimensions of this high-dimensional vector space. Hence we do not know how similar they are to what we think of as meaning or context. The model then takes these two vectors and produces a set of likelihoods for the next word; it selects and places one of the more likely ones—though not always the most likely. Allowing the model to choose randomly amongst the more likely words produces more creative and human-like text; the parameter which controls this is called the 'temperature' of the model and increasing the model's temperature makes it both seem more creative and more likely to produce falsehoods. The system then repeats the process until it has a recognizable, complete-looking response to whatever prompt it has been given.

Given this process, it's not surprising that LLMs have a problem with the truth. Their goal is to provide a normal-seeming response to a prompt, not to convey information that is helpful to their interlocutor. Examples of this are already numerous, for instance, a lawyer recently prepared his brief using ChatGPT and discovered to his chagrin that most of the cited cases were not real (Weiser, 2023);

as Judge P. Kevin Castel put it, ChatGPT produced a text filled with "bogus judicial decisions, with bogus quotes and bogus internal citations". Similarly, when computer science researchers tested ChatGPT's ability to assist in academic writing, they found that it was able to produce surprisingly comprehensive and sometimes even accurate text on biological subjects given the right prompts. But when asked to produce evidence for its claims, "it provided five references dating to the early 2000s. None of the provided paper titles existed, and all provided PubMed IDs (PMIDs) were of different unrelated papers" (Alkaissi and McFarland, 2023). These errors can "snowball": when the language model is asked to provide evidence for or a deeper explanation of a false claim, it rarely checks itself; instead it confidently producesmore false but normal-sounding claims (Zhang et al. 2023). The accuracy problem for LLMs and other generative Ais is often referred to as the problem of "AI hallucination": the chatbot seems to be hallucinating sources and facts that don't exist. These inaccuracies are referred to as "hallucinations" in both technical (OpenAI, 2023) and popular contexts (Weise & Metz, 2023).

These errors are pretty minor if the only point of a chatbot is to mimic human speech or communication. But the companies designing and using these bots have grander plans: chatbots could replace Google or Bing searches with a more user-friendly conversational interface (Shah & Bender, 2022; Zhu et al., 2023), or assist doctors or therapists in medical contexts (Lysandrou, 2023). In these cases, accuracy is important and the errors represent a serious problem.

One attempted solution is to hook the chatbot up to some sort of database, search engine, or computational program that can answer the questions that the LLM gets wrong (Zhu et al., 2023). Unfortunately, this doesn't work very well either. For example, when ChatGPT is connected to Wolfram Alpha, a powerful piece of mathematical software, it improves moderately in answering simple mathematical questions. But it still regularly gets things wrong, especially for questions which require multi-stage thinking (Davis & Aaronson, 2023). And when connected to search engines or other databases, the models are still fairly likely to provide fake information unless they are given very specific instructions–and even then things aren't perfect (Lysandrou, 2023). OpenAI has plans to rectify this by training the model to do step by step reasoning (Lightman et al., 2023) but this is quite resource-intensive, and there is reason to be doubtful that it will completely solve the problem—nor is it clear that the result will be a large language model, rather than some broader form of AI.

Solutions such as connecting the LLM to a database don't work is because, if the models are *trained* on the database, then the words in the database affect the probability that the chatbot will add one or another word to the line of text it is generating. But this will only make it produce text similar to the text in the database; doing so will make it more likely that it reproduces the information in the database but by no means ensures that it will.

On the other hand, the LLM can also be connected to the database by allowing it to consult the database, in a way similar to the way it consults or talks to its human interlocutors. In this way, it can use the outputs of the database as text which it responds to and builds on. Here's one way this can work: when a human interlocutor asks the language model a question, it can then translate the question into a query for the database. Then, it takes the response of the database as an input and builds a text from it to provide back to the human questioner. But this can misfire too, as the chatbots might ask the database the wrong question, or misinterpret its answer (Davis & Aaronson, 2023). "GPT-4 often struggles to formulate a problem in a way that Wolfram Alpha can accept or that produces useful output." This is not unrelated to the fact that when the language model generates a query for the database or computational module, it does so in the same way it generates text for humans: by estimating the likelihood that some output "looks like" the kind of thing the database will correspond with.

One might worry that these failed methods for improving the accuracy of chatbots are connected to the inapt metaphor of AI hallucinations. If the AI is *misperceiving* or *hallucinating* sources, one way to rectify this would be to put it in touch with real rather than hallucinated sources. But attempts to do so have failed.

The problem here isn't that large language models hallucinate, lie, or misrepresent the world in some way. It's that they are not designed to represent the world at all; instead, they are designed to convey convincing lines of text. So when they are provided with a database of some sort, they use this, in one way or another, to make their responses more convincing. But they are not in any real way attempting to convey or transmit the information in the database. As Chirag Shah and Emily Bender put it: "Nothing in the design of language models (whose training task is to predict words given context) is actually designed to handle arithmetic, temporal reasoning, etc. To the extent that they sometimes get the right answer to such questions is only because they happened to synthesize relevant strings out of what was in their training data. No reasoning is involved […] Similarly, language models are prone to making stuff up […] because they are not designed to express some underlying set of information in natural language; they are only manipulating the form of language" (Shah & Bender, 2022). These models aren't designed to transmit information, so we shouldn't be too surprised when their assertions turn out to be false.

Lies, 'hallucinations' and bullshit

Frankfurtian bullshit and lying

Many popular discussions of ChatGPT call its false statements 'hallucinations'. One also might think of these untruths as lies. However, we argue that this isn't the right way to think about it. We will argue that these falsehoods aren't hallucinations later – in Sect. 3.2.3. For now, we'll discuss why these untruths aren't lies but instead are bullshit.

The topic of lying has a rich philosophical literature. In 'Lying', Saint Augustine distinguished seven types of lies, and his view altered throughout his life. At one point, he defended the position that any instance of knowingly uttering a false utterance counts as a lie, so that even jokes containing false propositions, like –

> I entered a pun competition and because I really wanted to win, I submitted ten entries. I was sure one of them would win, but no pun in ten did.

– would be regarded as a lie, as I have never entered such a competition (Proops & Sorensen, 2023: 3). Later, this view is refined such that the speaker only lies if they intend the hearer to believe the utterance. The suggestion that the speaker must intend to deceive is a common stipulation in literature on lies. According to the "traditional account" of lying:

> To lie$=_{df.}$ to make a believed-false statement to another person with the intention that the other person believe that statement to be true (Mahon, 2015).

For our purposes this definition will suffice. Lies are generally frowned upon. But there are acts of misleading testimony which are criticisable, which do not fall under the umbrella of lying.[1] These include spreading untrue gossip, which one mistakenly, but culpably, believes to be true. Another class of misleading testimony that has received particular attention from philosophers is that of bullshit. This everyday notion was analysed and introduced into the philosophical lexicon by Harry Frankfurt.[2]

Frankfurt understands bullshit to be characterized not by an intent to deceive but instead by a reckless disregard for the truth. A student trying to sound knowledgeable without having done the reading, a political candidate saying things because they sound good to potential voters, and a dilettante trying to spin an interesting story: none of these people are trying to deceive, but they are also not trying to convey facts. To Frankfurt, they are bullshitting.

Like "lie", "bullshit" is both a noun and a verb: an utterance produced can be a lie or an instance of bullshit, as can the act of producing these utterances. For an utterance to be classed as bullshit, it must not be accompanied by the explicit intentions that one has when lying, i.e., to cause a false belief in the hearer. Of course, it must also not be accompanied by the intentions characterised by an honest utterance. So far this story is entirely negative. Must any positive intentions be manifested in the utterer?

Throughout most of Frankfurt's discussion, his characterisation of bullshit is negative. He notes that bullshit requires "no conviction" from the speaker about what the truth is (2005: 55), that the bullshitter "pays no attention" to the truth (2005: 61) and that they "may not deceive us, or even intend to do so, either about the facts or what he takes the facts to be" (2005: 54). Later, he describes the "defining feature" of bullshit as "*a lack of concern* with truth, or an *indifference to how things really are* [our emphasis]" (2002: 340). These suggest a negative picture; that for an output to be classed as bullshit, it only needs to lack a certain relationship to the truth.

However, in places, a positive intention is presented. Frankfurt says what a bullshitter ….

"…does necessarily attempt to deceive us about is his enterprise. His only indispensably distinctive characteristic is that in a certain way he misrepresents what he is up to" (2005: 54).

This is somewhat surprising. It restricts what counts as bullshit to utterances accompanied by a higher-order deception. However, some of Frankfurt's examples seem to lack this feature. When Fania Pascal describes her unwell state as "feeling like a dog that has just been run over" to her friend Wittgenstein, it stretches credulity to suggest that she was intending to deceive him about how much she knew about how run-over dogs felt. And given how the conditions for bullshit are typically described as negative, we might wonder whether the positive condition is really necessary.

Bullshit distinctions

Should utterances without an intention to deceive count as bullshit? One reason in favour of expanding the definition, or embracing a plurality of bullshit, is indicated by Frankfurt's comments on the dangers of bullshit.

"In contrast [to merely unintelligible discourse], indifference to the truth is extremely dangerous. The conduct of civilized life, and the vitality of the institutions that are

[1] A particularly surprising position is espoused by Fichte, who regards as lying not only lies of omission, but knowingly *not correcting* someone who is operating under a falsehood. For instance, if I was to wear a wig, and someone believed this to be my real hair, Fichte regards this as a lie, for which I am culpable. Bacin (2021) for further discussion of Fichte's position.

[2] Originally published in *Raritan*, VI(2) in 1986. References to that work here are from the 2005 book version.

indispensable to it, depend very fundamentally on respect for the distinction between the true and the false. Insofar as the authority of this distinction is undermined by the prevalence of bullshit and by the mindlessly frivolous attitude that accepts the proliferation of bullshit as innocuous, an indispensable human treasure is squandered" (2002: 343).

These dangers seem to manifest regardless of whether there is an intention to deceive about the enterprise a speaker is engaged in. Compare the deceptive bullshitter, who does aim to mislead us about being in the truth-business, with someone who harbours no such aim, but just talks for the sake of talking (without care, or indeed any thought, about the truth-values of their utterances).

One of Frankfurt's examples of bullshit seems better captured by the wider definition. He considers the advertising industry, which is "replete with instances of bullshit so unmitigated that they serve among the most indisputable and classic paradigms of the concept" (2005:22). However, it seems to misconstrue many advertisers to portray their aims as to mislead about their agendas. They are *expected* to say misleading things. Frankfurt discusses Marlboro adverts with the message that smokers are as brave as cowboys (2002: 341). Is it reasonable to suggest that the advertisers pretended to believe this?

Frankfurt does allow for multiple species of bullshit (2002: 340).[3] Following this suggestion, we propose to envisage bullshit as a genus, and Frankfurt's intentional bullshit as one species within this genus. Other species may include that produced by the advertiser, who anticipates that no one will believe their utterances[4] or someone who has no intention one way or another about whether they mislead their audience. To that end, consider the following distinction:

Bullshit (general) Any utterance produced where a speaker has indifference towards the truth of the utterance.

Hard bullshit Bullshit produced with the intention to mislead the audience about the utterer's agenda.

Soft bullshit Bullshit produced without the intention to mislead the hearer regarding the utterer's agenda.

The general notion of bullshit is useful: on some occasions, we might be confident that an utterance was either soft bullshit or hard bullshit, but be unclear which, given our ignorance of the speaker's higher-order desires.[5] In such a case, we can still call bullshit.

Frankfurt's own explicit account, with the positive requirements about producer's intentions, is hard bullshit, whereas soft bullshit seems to describe some of Frankfurt's examples, such as that of Pascal's conversation with Wittgenstein, or the work of advertising agencies. It might be helpful to situate these distinctions in the existing literature. On our view, hard bullshit is most closely aligned with Cassam (2019), and Frankfurt's positive account, for the reason that all of these views hold that some intention must be present, rather than merely absent, for the utterance to be bullshit: a kind of "epistemic insouciance" or vicious attitude towards truth on Cassam's view, and (as we have seen) an intent to mislead the hearer about the utterer's agenda on Frankfurt's view. In Sect. 3.2 we consider whether ChatGPT may be a hard bullshitter, but it is important to note that it seems to us that hard bullshit, like the two accounts cited here, requires one to take a stance on whether or not LLMs can be agents, and so comes with additional argumentative burdens.

Soft bullshit, by contrast, captures only Frankfurt's negative requirement – that is, the indifference towards truth that we have classed as definitional of bullshit (general) – for the reasons given above. As we argue, ChatGPT is at minimum a soft bullshitter or a bullshit machine, because if it is not an agent then it can neither hold any attitudes towards truth nor towards deceiving hearers about its (or, perhaps more properly, its users') agenda.

It's important to note that even this more modest kind of bullshitting will have the deleterious effects that concern Frankfurt: as he says, "indifference to the truth is extremely dangerous…by the mindlessly frivolous attitude

[3] In making this comment, Frankfurt concedes that what Cohen calls "bullshit" is also worthy of the name. In Cohen's use (2002), bullshit is a type of unclarifiable text, which he associates with French Marxists. Several other authors have also explored this area in various ways in recent years, each adding valuable nuggets to the debate. Dennis Whitcomb and Kenny Easwaran expand the domains to which "bullshit" can be applied. Whitcomb argues there can be bullshit questions (as well as propositions), whereas Easwaran argues that we can fruitfully view some activities as bullshit (2023).While we accept that these offer valuable streaks of bullshit insight, we will restrict our discussion to the Frankfurtian framework. For those who want to wade further into these distinctions, Neil Levy's Philosophy, Bullshit, and Peer Review (2023) offers a taxonomical overview of the bullshit out there.

[4] This need not undermine their goal. The advertiser may intend to impress associations (e.g., positive thoughts like "cowboys" or "brave" with their cigarette brand) upon their audience, or reinforce/instil brand recognition.Frankfurt describes this kind of scenario as occurring in a "bull session": "Each of the contributors to a bull session relies…upon a general recognition that what he expresses or says is not to be understood as being what he means wholeheartedly or believes unequivocally to be true" (2005: 37). Yet Frankfurt claims that the contents of bull sessions are distinct from bullshit.

[5] It's worth noting that something like the distinction between hard and soft bullshitting we draw also occurs in Cohen (2002): he suggests that we might think of someone as a bullshitter as "a person who aims at bullshit, however frequently or infrequently he hits his target", *or* if they are merely "disposed to bullshit: for whatever reason, to produce a lot of unclarifiable stuff" (p334). While we do not adopt Cohen's account here, the parallels between his characterisation and our own are striking.

that accepts the proliferation of bullshit as innocuous, an indispensable human treasure is squandered" (2002, p343). By treating ChatGPT and similar LLMs as being in any way concerned with truth, or by speaking metaphorically as if they make mistakes or suffer "hallucinations" in pursuit of true claims, we risk exactly this acceptance of bullshit, and this squandering of meaning – so, irrespective of whether or not ChatGPT is a hard or a soft bullshitter, it does produce bullshit, and it does matter.

ChatGPT is bullshit

With this distinction in hand, we're now in a position to consider a worry of the following sort: Is ChatGPT hard bullshitting, soft bullshitting, or neither? We will argue, first, that ChatGPT, and other LLMs, are clearly soft bullshitting. However, the question of whether these chatbots are hard bullshitting is a trickier one, and depends on a number of complex questions concerning whether ChatGPT can be ascribed intentions. We canvas a few ways in which ChatGPT can be understood to have the requisite intentions in Sect. 3.2.

ChatGPT is a soft bullshitter

We are not confident that chatbots can be correctly described as having any intentions at all, and we'll go into this in more depth in the next Sect. (3.2). But we are quite certain that ChatGPT does not intend to convey truths, and so is a soft bullshitter. We can produce an easy argument by cases for this. Either ChatGPT has intentions or it doesn't. If ChatGPT has no intentions at all, it trivially doesn't intend to convey truths. So, it is indifferent to the truth value of its utterances and so is a soft bullshitter.

What if ChatGPT does have intentions? In Sect. 1, we argued that ChatGPT is not designed to produce true utterances; rather, it is designed to produce text which is indistinguishable from the text produced by humans. It is aimed at being convincing rather than accurate. The basic architecture of these models reveals this: they are designed to come up with a *likely continuation of a string of text.* It's reasonable to assume that one way of being a likely continuation of a text is by being true; if humans are roughly more accurate than chance, true sentences will be more likely than false ones. This might make the chatbot more accurate than chance, but it does not give the chatbot any intention to convey truths. This is similar to standard cases of human bullshitters, who don't care whether their utterances are true; good bullshit often contains some degree of truth, that's part of what makes it convincing. A bullshitter can be more accurate than chance while still being indifferent to the truth of their utterances. We conclude that, even if the chatbot can be described as having intentions, it is indifferent to whether its utterances are true. It does not and cannot care about the truth of its output.

Presumably ChatGPT can't care about conveying or hiding the truth, since it can't care about anything. So, just as a matter of conceptual necessity, it meets one of Frankfurt's criteria for bullshit. However, this only gets us so far – a rock can't care about anything either, and it would be patently absurd to suggest that this means rocks are bullshitters[6]. Similarly books can contain bullshit, but they are not themselves bullshitters. Unlike rocks – or even books – ChatGPT itself produces text, and looks like it performs speech acts independently of its users and designers. And while there is considerable disagreement concerning whether ChatGPT has intentions, it's widely agreed that the sentences it produces are (typically) meaningful (see e.g. Mandelkern and Linzen 2023).

ChatGPT functions not to convey truth or falsehood but rather to convince the reader of – to use Colbert's apt coinage – the *truthiness* of its statement, and ChatGPT is designed in such a way as to make attempts at bullshit efficacious (in a way that pens, dictionaries, etc., are not). So, it seems that at minimum, ChatGPT is a soft bullshitter: if we take it not to have intentions, there isn't any attempt to mislead about the attitude towards truth, but it *is* nonetheless engaged in the business of outputting utterances that look as if they're truth-apt. We conclude that ChatGPT is a *soft bullshitter.*

ChatGPT as hard bullshit

But is ChatGPT a *hard bullshitter*? A critic might object, it is simply inappropriate to think of programs like ChatGPT as hard bullshitters, because (i) they are not agents, or relatedly, (ii) they do not and cannot intend anything whatsoever.

We think this is too fast. First, whether or not ChatGPT has agency, its creators and users do. And what they produce with it, we will argue, is bullshit. Second, we will argue that, regardless of whether it has agency, it does have a function; this function gives it characteristic goals, and possibly even intentions, which align with our definition of hard bullshit.

Before moving on, we should say what we mean when we ask whether ChatGPT is an agent. For the purposes of

[6] Of course, rocks also can't express propositions – but then, part of the worry here is whether ChatGPT actually *is* expressing propositions, or is simply a means through which agents express propositions. A further worry is that we shouldn't even see ChatGPT as expressing propositions - perhaps there are no communicative intentions, and so we should see the outputs as meaningless. Even accepting this, we can still meaningfully talk about them as expressing propositions. This proposal - fictionalism about chatbots - has recently been discussed by Mallory (2023).

this paper, the central question is whether ChatGPT has intentions and or beliefs. Does it intend to deceive? Can it, in any literal sense, be said to have goals or aims? If so, does it intend to deceive us about the content of its utterances, or merely have the goal to appear to be a competent speaker? Does it have beliefs—internal representational states which aim to track the truth? If so, do its utterances match those beliefs (in which case its false statements might be something like hallucinations) or are its utterances not matched to the beliefs—in which case they are likely to be either lies or bullshit? We will consider these questions in more depth in Sect. 3.2.2.

There are other philosophically important aspects of agenthood that we will not be considering. We won't be considering whether ChatGPT makes decisions, has or lacks autonomy, or is conscious; we also won't worry whether ChatGPT is morally responsible for its statements or its actions (if it has any of those).

ChatGPT is a bullshit machine

We will argue that even if ChatGPT is not, itself, a hard bullshitter, it is nonetheless a bullshit machine. The bullshitter is the person using it, since they (i) don't care about the truth of what it says, (ii) want the reader to believe what the application outputs. On Frankfurt's view, bullshit is bullshit even if uttered with no intent to bullshit: if something is bullshit to start with, then its repetition "is bullshit as he [or it] repeats it, insofar as it was originated by someone who was unconcerned with whether what he was saying is true or false" (2022, p340).

This just pushes the question back to who the originator is, though: take the (increasingly frequent) example of the student essay created by ChatGPT. If the student cared about accuracy and truth, they would not use a program that infamously makes up sources whole-cloth. Equally, though, if they give it a prompt to produce an essay on philosophy of science and it produces a recipe for Bakewell tarts, then it won't have the desired effect. So the idea of ChatGPT as a bullshit machine seems right, but also as if it's missing something: someone can produce bullshit using their voice, a pen or a word processor, after all, but we don't standardly think of these things as being bullshit machines, or of outputting bullshit in any particularly interesting way – conversely, there *does* seem to be something particular to ChatGPT, to do with the way that it operates, which makes it more than a *mere* tool, and which suggests that it might appropriately be thought of as an originator of bullshit. In short, it doesn't seem quite right either to think of ChatGPT as analogous to a pen (can be used for bullshit, but can create nothing without deliberate and wholly agent-directed action) nor as to a bullshitting human (who can intend and produce bullshit on their own initiative).

The idea of ChatGPT as a bullshit machine is a helpful one when combined with the distinction between hard and soft bullshit. Reaching again for the example of the dodgy student paper: we've all, I take it, marked papers where it was obvious that a dictionary or thesaurus had been deployed with a crushing lack of subtlety; where fifty-dollar words are used not because they're the best choice, nor even because they serve to obfuscate the truth, but simply because the author wants to convey an *impression* of understanding and sophistication. It would be inappropriate to call the dictionary a bullshit artist in this case; but it would *not* be inappropriate to call the result bullshit. So perhaps we should, strictly, say not that ChatGPT *is* bullshit but that it *outputs* bullshit in a way that goes beyond being simply a vector of bullshit: it does not and cannot care about the truth of its output, *and* the person using it does so not to convey truth or falsehood but rather to convince the hearer that the text was written by a interested and attentive agent.

ChatGPT may be a hard bullshitter

Is ChatGPT itself a hard bullshitter? If so, it must have intentions or goals: it must intend to deceive its listener, not about the content of its statements, but instead about its agenda. Recall that hard bullshitters, like the unprepared student or the incompetent politician, don't care whether their statements are true or false, but do intend to deceive their audience about what they are doing. If so, it must have intentions or goals: it must intend to deceive its listener, not about the content of its statements, but instead about its agenda. We don't think that ChatGPT is an agent or has intentions in precisely the same way that humans do (see Levenstein and Herrmann (forthcoming) for a discussion of the issues here). But when speaking loosely it is remarkably easy to use intentional language to describe it: what is ChatGPT *trying* to do? Does it *care* whether the text it produces is accurate? We will argue that there is a robust, although perhaps not literal, sense in which ChatGPT does intend to deceive us about its agenda: its goal is not to convince us of the content of its utterances, but instead to portray itself as a 'normal' interlocutor like ourselves. By contrast, there is no similarly strong sense in which ChatGPT confabulates, lies, or hallucinates.

Our case will be simple: ChatGPT's primary function is to imitate human speech. If this function is intentional, it is precisely the sort of intention that is required for an agent to be a hard bullshitter: in performing the function, ChatGPT is attempting to deceive the audience about its agenda. Specifically, it's trying to seem like something that has an agenda, when in many cases it does not. We'll discuss here

whether this function gives rise to, or is best thought of, as an intention. In the next Sect. (3.2.3), we will argue that ChatGPT has no similar function or intention which would justify calling it a confabulator, liar, or hallucinator.

How do we know that ChatGPT functions as a hard bullshitter? Programs like ChatGPT are designed to do a task, and this task is remarkably like what Frankfurt thinks the bullshitter intends, namely to deceive the reader about the nature of the enterprise – in this case, to deceive the reader into thinking that they're reading something produced by a being with intentions and beliefs.

ChatGPT's text production algorithm was developed and honed in a process quite similar to artificial selection. Functions and selection processes have the same sort of directedness that human intentions do; naturalistic philosophers of mind have long connected them to the intentionality of human and animal mental states. If ChatGPT is understood as having intentions or intention-like states in this way, its intention is to present itself in a certain way (as a conversational agent or interlocutor) rather than to represent and convey facts. In other words, it has the intentions we associate with hard bullshitting.

One way we can think of ChatGPT as having intentions is by adopting Dennett's *intentional stance* towards it. Dennett (1987: 17) describes the intentional stance as a way of predicting the behaviour of systems whose purpose we don't already know.

"To adopt the intentional stance […] is to decide – tentatively, of course – to attempt to characterize, predict, and explain […] behavior by using intentional idioms, such as 'believes' and 'wants,' a practice that assumes or presupposes the rationality" of the target system (Dennett, 1983: 345).

Dennett suggests that if we know why a system was designed, we can make predictions on the basis of its design (1987). While we do know that ChatGPT was designed to chat, its exact algorithm and the way it produces its responses has been developed by machine learning, so we do not know its precise details of how it works and what it does. Under this ignorance it is tempting to bring in intentional descriptions to help us understand and predict what ChatGPT is doing.

When we adopt the intentional stance, we will be making bad predictions if we attribute any desire to convey truth to ChatGPT. Similarly, attributing "hallucinations" to ChatGPT will lead us to predict as if it has perceived things that aren't there, when what it is doing is much more akin to making something up because it sounds about right. The former intentional attribution will lead us to try to correct its beliefs, and fix its inputs --- a strategy which has had limited if any success. On the other hand, if we attribute to ChatGPT the intentions of a hard bullshitter, we will be better able to diagnose the situations in which it will make mistakes and convey falsehoods. If ChatGPT is trying to do anything, it is trying to portray itself as a person.

Since this reason for thinking ChatGPT is a hard bullshitter involves committing to one or more controversial views on mind and meaning, it is more tendentious than simply thinking of it as a bullshit machine; but regardless of whether or not the program has intentions, there clearly *is* an attempt to deceive the hearer or reader about the nature of the enterprise somewhere along the line, and in our view that justifies calling the output hard bullshit.

So, though it's worth making the caveat, it doesn't seem to us that it significantly affects how we should think of and talk about ChatGPT and bullshit: the person using it to turn out some paper or talk isn't concerned either with conveying or covering up the truth (since both of those require attention to what the truth actually *is*), and neither is the system itself. Minimally, it churns out soft bullshit, and, given certain controversial assumptions about the nature of intentional ascription, it produces hard bullshit; the specific texture of the bullshit is not, for our purposes, important: either way, ChatGPT is a bullshitter.

Bullshit? hallucinations? confabulations? The need for new terminology

We have argued that we should use the terminology of bullshit, rather than "hallucinations" to describe the utterances produced by ChatGPT. The suggestion that "hallucination" terminology is inappropriate has also been noted by Edwards (2023), who favours the term "confabulation" instead. Why is our proposal better than this or other alternatives?

We object to the term hallucination because it carries certain misleading implications. When someone hallucinates they have a non-standard perceptual experience, but do not actually perceive some feature of the world (Macpherson, 2013), where "perceive" is understood as a success term, such that they do not actually perceive the object or property. This term is inappropriate for LLMs for a variety of reasons. First, as Edwards (2023) points out, the term hallucination anthropomorphises the LLMs. Edwards also notes that attributing resulting problems to "hallucinations" of the models may allow creators to "blame the AI model for faulty outputs instead of taking responsibility for the outputs themselves", and we may be wary of such abdications of responsibility. LLMs do not perceive, so they surely do not "mis-perceive". Second, what occurs in the case of an LLM delivering false utterances is not an unusual or deviant form of the process it usually goes through (as some claim is the case in hallucinations, e.g., disjunctivists about

perception). The very same process occurs when its outputs happen to be true.

So much for "hallucinations". What about Edwards' preferred term, "confabulation"? Edwards (2023) says:

> In human psychology, a "confabulation" occurs when someone's memory has a gap and the brain convincingly fills in the rest without intending to deceive others. ChatGPT does not work like the human brain, but the term "confabulation" arguably serves as a better metaphor because there's a creative gap-filling principle at work […].

As Edwards notes, this is imperfect. Once again, the use of a human psychological term risks anthropomorphising the LLMs.

This term also suggests that there is something exceptional occurring when the LLM makes a false utterance, i.e., that in these occasions - and only these occasions - it "fills in" a gap in memory with something false. This too is misleading. Even when the ChatGPT does give us correct answers, its process is one of predicting the next token. In our view, it falsely indicates that ChatGPT is, in general, attempting to convey accurate information in its utterances. But there are strong reasons to think that it does not have beliefs that it is intending to share in general–see, for example, Levenstein and Herrmann (forthcoming). In our view, it falsely indicates that ChatGPT is, in general, attempting to convey accurate information in its utterances. Where it does track truth, it does so indirectly, and incidentally.

This is why we favour characterising ChatGPT as a bullshit machine. This terminology avoids the implications that perceiving or remembering is going on in the workings of the LLM. We can also describe it as bullshitting whenever it produces outputs. Like the human bullshitter, some of the outputs will likely be true, while others not. And as with the human bullshitter, we should be wary of relying upon any of these outputs.

Conclusion

Investors, policymakers, and members of the general public make decisions on how to treat these machines and how to react to them based not on a deep technical understanding of how they work, but on the often metaphorical way in which their abilities and function are communicated. Calling their mistakes 'hallucinations' isn't harmless: it lends itself to the confusion that the machines are in some way *misperceiving* but are nonetheless trying to convey something that they believe or have perceived. This, as we've argued, is the wrong metaphor. The machines are not trying to communicate something they believe or perceive. Their inaccuracy is not due to misperception or hallucination. As we have pointed out, they are not trying to convey information at all. They are bullshitting.

Calling chatbot inaccuracies 'hallucinations' feeds in to overblown hype about their abilities among technology cheerleaders, and could lead to unnecessary consternation among the general public. It also suggests solutions to the inaccuracy problems which might not work, and could lead to misguided efforts at AI alignment amongst specialists. It can also lead to the wrong attitude towards the machine when it gets things right: the inaccuracies show that it is bullshitting, even when it's right. Calling these inaccuracies 'bullshit' rather than 'hallucinations' isn't just more accurate (as we've argued); it's good science and technology communication in an area that sorely needs it.

Acknowledgements Thanks to Neil McDonnell, Bryan Pickel, Fenner Tanswell, and the University of Glasgow's Large Language Model reading group for helpful discussion and comments.

References

Alkaissi, H., & McFarlane, S. I., (2023, February 19). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus, 15*(2), e35179. https://doi.org/10.7759/cureus.35179.

Bacin, S. (2021). My duties and the morality of others: Lying, truth and the good example in Fichte's normative perfectionism. In S. Bacin, & O. Ware (Eds.), *Fichte's system of Ethics: A critical guide*. Cambridge University Press.

Cassam, Q. (2019). *Vices of the mind*. Oxford University Press.

Cohen, G. A. (2002). Deeper into bullshit. In S. Buss, & L. Overton (Eds.), *The contours of Agency: Essays on themes from Harry Frankfurt*. MIT Press.

Davis, E., & Aaronson, S. (2023). Testing GPT-4 with Wolfram alpha and code interpreter plub-ins on math and science problems. *Arxiv Preprint: arXiv, 2308*, 05713v2.

Dennett, D. C. (1983). Intentional systems in cognitive ethology: The panglossian paradigm defended. *Behavioral and Brain Sciences*, *6*, 343–390.

Dennett, D. C. (1987). *The intentional stance*. The MIT.

Dennis Whitcomb (2023). Bullshit questions. *Analysis*, *83*(2), 299–304.

Easwaran, K. (2023). Bullshit activities. *Analytic Philosophy*, *00*, 1–23. https://doi.org/10.1111/phib.12328.

Edwards, B. (2023). Why ChatGPT and bing chat are so good at making things up. *Ars Tecnica*. https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/, accesssed 19th April, 2024.

Frankfurt, H. (2002). Reply to cohen. In S. Buss, & L. Overton (Eds.), *The contours of agency: Essays on themes from Harry Frankfurt*. MIT Press.

Frankfurt, H. (2005). *On Bullshit*, Princeton.

Knight, W. (2023). Some glimpse AGI in ChatGPT. others call it a mirage. *Wired*, August 18 2023, accessed via https://www.wired.com/story/chatgpt-agi-intelligence/.

Levenstein, B. A., & Herrmann, D. A. (forthcoming). Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, 1–27.

Levy, N. (2023). *Philosophy, Bullshit, and peer review*. Camridge University.

Lightman, H., et al. (2023). Let's verify step by step. *Arxiv Preprint: arXiv*, *2305*, 20050.

Lysandrou (2023). Comparative analysis of drug-GPT and ChatGPT LLMs for healthcare insights: Evaluating accuracy and relevance in patient and HCP contexts. *ArXiv Preprint: arXiv*, *2307*, 16850v1.

Macpherson, F. (2013). The philosophy and psychology of hallucination: an introduction, in *Hallucination*, Macpherson and Platchias (Eds.), London: MIT Press.

Mahon, J. E. (2015). The definition of lying and deception. *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (Ed.), https://plato.stanford.edu/archives/win2016/entries/lying-definition/.

Mallory, F. (2023). Fictionalism about chatbots. *Ergo*, *10*(38), 1082–1100.

Mandelkern, M., & Linzen, T. (2023). Do language models' Words Refer?. *ArXiv Preprint: arXiv*, *2308*, 05576.

OpenAI (2023). GPT-4 technical report. *ArXiv Preprint: arXiv*, *2303*, 08774v3.

Proops, I., & Sorensen, R. (2023). Destigmatizing the exegetical attribution of lies: the case of kant. *Pacific Philosophical Quarterly*. https://doi.org/10.1111/papq.12442.

Sarkar, A. (2023). ChatGPT 5 is on track to attain artificial general intelligence. *The Statesman*, April 12, 2023. Accesses via https://www.thestatesman.com/supplements/science_supplements/chatgpt-5-is-on-track-to-attain-artificial-general-intelligence-1503171366.html.

Shah, C., & Bender, E. M. (2022). Situating search. CHIIR '22: Proceedings of the 2022 Conference on Human Information Interaction and Retrieval March 2022 Pages 221–232 https://doi.org/10.1145/3498366.3505816.

Weise, K., & Metz, C. (2023). When AI chatbots hallucinate. New York Times, May 9, 2023. Accessed via https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html.

Weiser, B. (2023). Here's what happens when your lawyer uses ChatGPT. *New York Times*, May 23, 2023. Accessed via https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html.

Zhang (2023). How language model hallucinations can snowball. *ArXiv preprint: arXiv:*, *2305*, 13534v1.

Zhu, T., et al. (2023). Large language models for information retrieval: A survey. *Arxiv Preprint: arXiv*, *2308*, 17107v2.