How Not to Talk about Chatbot Mistakes

Preprint before Review, 6 September 2024

Markus Pantsar*
Human Technology Center, RWTH Aachen University
markus.pantsar@humtec.rwth-aachen.de
ORCID: 0000-0001-8572-1453

Regina E. Fabry*
Department of Philosophy, Macquarie University
regina.fabry@mq.edu.au
ORCID: 0000-0003-1078-1499

*Both authors contributed equally to this work.

How Not to Talk about Chatbot Mistakes

Markus Pantsar & Regina E. Fabry

Abstract. The function of chatbots like OpenAI's ChatGPT is based on detecting probabilistic patterns in the training data. This makes them vulnerable to generating factual mistakes in their outputs. Recently, it has become commonplace in philosophical, scientific, and popular discourses to capture such mistakes by metaphors that draw on discourses about the human mind. The three most popular metaphors at present are hallucinating, confabulating, and bullshitting. In this paper, we review, discuss, and criticise these mental metaphors. By applying conceptual metaphor theory, we provide numerous reasons why none of the metaphors succeed in providing us with a better understanding of factual chatbot mistakes. We conclude by calling for justifications of the epistemic feasibility and fruitfulness of the metaphors at issue. Furthermore, we raise the question what would be lost if we stopped trying to capture factual chatbot mistakes by mental metaphors.

Keywords. ChatGPT; chatbot mistakes; AI hallucinations; conceptual metaphor theory; mental metaphors

1. Introduction

When OpenAI released the chatbot ChatGPT in late November 2022, scholars, scientists, teachers, journalists, and the wider public reacted with an intriguing combination of excitement and concern. ChatGPT sparked excitement, as it provided human agents with genuinely novel opportunities to communicate with a chatbot that could generate coherent and cohesive texts on virtually any topic in response to a simple prompt. In many cases, the textual outputs seem indistinguishable from those by human agents, which enables them to be used to augment or even replace humanly produced texts. In turn, ChatGPT gave rise to deep concerns, not least because it challenged socio-culturally shaped practices of written communication, be it at the corporate workplace, in schools and universities, or in the writer's studio. In addition to such concerns, the early excitement has also cooled down due to ChatGPT's limitations. As more and more agents engaged with it, one important limitation has become obvious: ChatGPT frequently generates outputs that are factually incorrect. These factual mistakes pose a significant limitation on the reliability of ChatGPT outputs across a wide range of domains, from history to geography and mathematics.

Philosophers, AI researchers and developers, popular science writers, and journalists were quick to publish articles and opinion pieces about ChatGPT's tendency to generate factually incorrect output. To capture and characterise these chatbot mistakes, they have often relied on metaphors that originate in discourses about the human mind. Specifically, certain aspects of human perception, action, and cognition have been used as a source domain for capturing chatbot mistakes in metaphorical terms. In what follows, we will refer to instances of this kind of conceptual resource as *mental metaphors*.

At the time of writing, the most frequently used and discussed mental metaphors to capture chatbot mistakes are hallucination (e.g., Heersmink et al., 2024; OpenAI et al., 2024; Weise & Metz, 2023), confabulation (e.g., Edwards, 2023; Henriques, 2024), and bullshitting (e.g., Bergstrom & Ogbundu, 2023; Hicks et al., 2024). However, in this paper, we will demonstrate that none of these metaphors manage to provide us with a better understanding of factual chatbot mistakes. We will argue that ChatGPT – and other chatbots such as Google's Gemini by extension – do not hallucinate, confabulate, or bullshit when they generate factually incorrect outputs, not even in a metaphorical sense. Drawing on research on hallucination, confabulation, and bullshitting in philosophy and the empirical cognitive sciences, we will show that the human mind is not an appropriate source domain for capturing chatbot mistakes. In developing this argument, the paper unfolds as follows. We first describe the workings of chatbots such as ChatGPT (Section 2) and then characterise chatbot mistakes (Section 3). To be in a position to analyse the mental metaphors that have been used to capture chatbot mistakes, we introduce conceptual metaphor theory (Lakoff & Johnson, 2003) as a framework and identify three ways in which metaphors can fail to be conducive to a better understanding of the target phenomenon (Section 4). We will then review, discuss, and criticise the hallucination, confabulation, and bullshitting metaphors that have been widely used in scholarly, scientific, technological, and popular discourses (Section 5). We end the paper with a discussion and systematisation of the outcomes of our analyses and respond to two possible objections (Section 6).

2. How Do Chatbots Work?

The history of chatbots can be traced back to Turing's (1950) "Imitation Game", a test for machine intelligence based on its ability to imitate human (textual) conversational behaviour. For Turing, the test was a theoretical notion – perhaps even meant to be a thought experiment, rather than an actual experiment to be run (Gonçalves, 2023) – but it gave the spark to many concrete developments in natural language communication between humans and computers. For a long time, the most famous of these developments was ELIZA, a simple program that responded based on the presence of keywords in the prompt – or in absence of a keyword, with a content-free remark (Weizenbaum, 1966). In the most famous incarnation of ELIZA, it mimicked a Rogerian therapist (Bassett, 2019). ELIZA, while successful in fooling many people into believing it was human, was a very simple rule-based program. In the following decades, more complex such programs were introduced, including the role-playing "Chatter Bot", which gave the name – soon to be abbreviated to "chatbot" – to natural language conversational programs (Mauldin, 1994).

While ELIZA and Chatter Bot implemented the fundamental principle of chatbots – namely that they generated text based on prompts – the modern notion of a 'chatbot' has come to refer to online generative AI systems run on *transformer*-based deep artificial neural networks. In the transformer architecture, pieces of text are converted to numerical representations (called "tokens"), and each token is contextualized as a vector through a self-attention mechanism to determine its importance (Vaswani et al., 2017). This methodology allows for the unsupervised pre-training of the system, which makes it possible to train it on vast amounts of data. For this reason, the technology is particularly suitable for training unimodal and multimodal *large language models* (LLMs).[1] In training such models, the dataset of tokens is run through an auto-regressive decoder that then ranks which token most probably follows a particular sequence of tokens. Hence, the pre-training of LLMs is based essentially on detecting *patterns* in the training data. Importantly, the architecture allows for *few-shot learning*, meaning that the LLM – using its vast general training material – can "learn" to make accurate predictions in new domains based on a relatively small set of training data (T. Brown et al., 2020).

At present, the most famous LLMs and multimodal LLMs are OpenAI's GPT (Generative Pre-trained Transformer) models. The amount of data used to train these models is enormous. GPT-2 had 1.5 billion parameters and was trained on a dataset of 8 million webpages. GPT-3 already had 175 billion parameters and was trained on 45 terabytes of data. The multimodal GPT-4 is said to be trained on 1 *peta*byte of data and has roughly 1.7 *trillion* parameters (T. Brown et al., 2020).[2] The GPT models can be used for different types of natural language processing tasks, but for most people, they are known through their application in the online chatbot ChatGPT. The functioning of ChatGPT is based on the probabilistic structure of the GPT model. When the user enters a

[1] A unimodal LLM processes only textual input while a multimodal LLM (sometimes abbreviated as MLLM) can include other media, such as images, audio, and video.

[2] The numbers for GPT-4 have not been officially released, but they have been widely circulated online. See: https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/

prompt into ChatGPT, the chatbot accesses the model to determine which tokens are most probably associated with the input tokens, and then generates an output of those tokens as natural language text. As users of ChatGPT know, the chat events are path-dependent in that the chatbot's responses to new prompts are determined also by the previous prompts (and responses) in the event. This helps give users the experience that they are really engaged in a dialogue, rather than merely consulting an online information resource, like Wikipedia. This impression can also be aided by the fact that ChatGPT does not always choose the token with the highest probability: to simulate the variety of human writing better, the GPT model can also be used to choose randomly from the tokens with the highest probabilities.

ChatGPT is certainly not the only chatbot on the market, but it is currently by far both the most popular and most discussed one (e.g., Heersmink et al., 2024; Hicks et al., 2024). Hence, in this paper we focus on it, treating ChatGPT as a representative of modern chatbot technology that is widely used. While the latest GPT model is a multimodal LLM, we will focus only on textual inputs and outputs to keep the topic manageable. All our considerations can be, *mutatis mutandis*, extended to multimodal models.

3. What Is a Chatbot Mistake?

For ChatGPT, the pre-trained GPT large language model is fine-tuned through human feedback (Ouyang et al., 2022). This allows it to give more useful and accurate answers to its users, as well as to avoid material that some users could consider harmful (Deshpande et al., 2023). The fine-tuning is therefore a crucial step in developing the chatbot outputs to be more desirable to its users (and consequently its developers, too). Given the aims of fine-tuning, some criteria for desirability of the outputs can be identified. First, the chatbot is supposed to be successful in its natural language processing, i.e., the generated outputs ought to be instances of grammatically correct and coherent, comprehensible language. Second, the chatbot is supposed to be accurate in its answers, i.e., the outputs should be factually correct. Third, the chatbot should not generate harmful or otherwise undesired types of output. For example, it should not be biased against members of certain groups (e.g., women, people of colour) (Bender et al., 2021; Kasirzadeh & Gabriel, 2023).

These three general aims of the outputs can be used to analyse what happens when ChatGPT makes a mistake in its output. First, it can give an ungrammatical, incoherent, or otherwise erroneous output on a linguistic level. Second, it can give a factually incorrect output. Third, it can produce an output that is offensive, discriminatory, or otherwise against the ethical aims stated by OpenAI. In this paper, we will focus only on the second type of a mistake, in which ChatGPT presents something as a fact in a mistaken way. To capture this phenomenon, we need to understand how the chatbot applies the LLM in generating the output. But before we move on to that topic, it is necessary to briefly discuss the issue of computer errors in a wider context. Factually incorrect ChatGPT outputs belong to a particular class of computer mistakes, and it is important not to confuse them with other types of errors.

Computer errors can be divided roughly into hardware and software defects. Hardware defects can be independent of the software being run or they can be connected to the software (e.g., through excessive load on the hardware). This latter case can be considered as one type of software defect. In addition to that, there are many different types of software defects. Typically, the notion of ‘error’ refers to cases in which there is a more general flaw in the development of a software. ’Bug’ is used when the software does not run in the way the programmers intended. Under this distinction, bugs can be thought of as manifestations of software errors (Sheremeta, 2023). Another, wider notion that is often used is a software ’issue’, which refers to some problem involved in the development and use of software (Bose, 2023).

Against this background, what happens when ChatGPT produces the second type of mistake identified above, i.e., when it presents something as a fact that is mistaken? Typically, we assume, the issue is unrelated to either hardware defects or bugs in the software. While such issues may have caused problems in the development of ChatGPT, they are unlikely to cause that kind of mistake in the output. It is also unlikely that hardware defects or bugs during the pre-training are responsible for the mistakes. Instead, it is likely that the pre-training process and the fine-tuning have not been successful in detecting the desired patterns in the dataset that are needed for factually accurate responses. By “desired”, we here mean the patterns that the developers identify as the ones most likely to be satisfactory to chatbot users. Output based on GPT detecting undesirable patterns could be called an error, but considering the design of LLMs, that would be misleading. Given the vast amount of data used for pre-training, and limited resources for fine-tuning, it seems impossible that ChatGPT would make no factual mistakes. The very strengths of the GPT large language model, from general pattern detection to few-shot learning, imply that it is a system designed to work most of the time, i.e., to provide correct and relevant responses to most prompts. When there is content that is not easily incorporated into the model in terms of its fit with more general patterns, mistakes are bound to happen. Hence mistakes of the second type as specified above (as well as the first and third types) are perhaps more accurately described as *issues* in ChatGPT.

Referring to ChatGPT mistakes as issues seems to be consistent with the consensual approach to discussing ChatGPT mistakes, even though the matter is often discussed in terms of errors by scientists, philosophers, and the media. Instead of trying to make ChatGPT free of mistakes, which is seen as an unrealistic pursuit, the focus has been on mitigating its mistakes (see, e.g., Biswas, 2023). This seems like a reasonable attitude. If ChatGPT is a useful tool, it would be unwise (and financially unviable) to make its use conditional on mistake-free outputs. Instead, we should aim to optimize the use of ChatGPT (and other chatbots). In addition to developing LLMs further and fine-tuning their applications for chatbots, it is important to develop a better understanding of the kind of mistakes that ChatGPT typically makes, why it makes them, and how they can be detected. Therefore, dealing with chatbot mistakes should also be (partly) the responsibility of its users. But in order to do that, users need some kind of awareness of what is actually happening in the processing of the software. Here we contend that in order to educate users to that effect, it is important to characterise the functioning – including the mistakes – of ChatGPT without using misleading or confusing terminology. Unfortunately, much of the reporting and philosophical theorising on the issue of ChatGPT mistakes does exactly that, calling what the chatbot is doing "hallucinating" (e.g., Heersmink et al., 2024; Maynez et al., 2020; OpenAI et al., 2024; Weise & Metz, 2023), "confabulating" (e.g., Edwards, 2023; Henriques, 2024; Rodgers et al., 2023), or "bullshitting" (e.g., Bergstrom & Ogbundu, 2023; Hicks et al., 2024).

Granted, few if any of the authors using such terminology are likely to think that what ChatGPT is doing is literally similar to what humans are doing when they are hallucinating, confabulating, or bullshitting. Clearly, these notions are used metaphorically. But do they stand scrutiny as successful metaphors? To answer this question, we will first consider conceptual metaphor theory (Section 4) and then discuss in detail the widely used hallucination, confabulation, and bullshitting metaphors (Section 5).

4. Conceptual Metaphor Theory

To be in a position to describe and evaluate the different uses of metaphorical language to discuss chatbot mistakes, we first need to better understand what metaphors are, how they work, and how they shape our thinking across scholarly and public discourses. To this end, we use *conceptual metaphor theory* introduced by Lakoff and Johnson (1980/2003) as a framework.[3] How do conceptual metaphors work? The basic setting of a metaphor includes a *source domain* and a *target domain*, as well as a mapping between the two (Lakoff & Johnson, 2003). The mapping takes items from the source domain and maps them to the target domain. The metaphor is successful if, by grasping the items and processes in the source domain, we can better grasp the mapped items and processes in the target domain. Lakoff and Johnson (2003) and Fernandez-Duque and Johnson (1999) use the mind-as-machine metaphor as an example. Here machine is the source domain and our grasp of machines (e.g., functions, products, automated functioning) can help us understand the mind (in this case, mental capacities, ideas and thinking, respectively). If the source domain of machines is sufficiently similar to the target domain, and if the conceptual mapping between the two is apt, then the mind-as-machine metaphor makes sense and helps us better understand the target phenomenon. The mapping then "gives rise to a systematic use of ordinary, conventional linguistic expressions in much of our talk about mental operations" (Fernandez-Duque & Johnson, 1999, p. 85). Hence, the metaphor enables the use of expressions like "I'm a little rusty today" and "we're running out of steam" to be informative about mental phenomena (Lakoff & Johnson, 2003, p. 27).

The key idea behind conceptual metaphor theory was that metaphors are not merely a linguistic phenomenon, but a crucial part of human cognition. Thus, conceptual metaphor theory has been applied to phenomena like gestures (Lakoff & Johnson, 2003), attention (Fernandez-Duque & Johnson, 1999), and mathematical cognition (Lakoff & Núñez, 2000; Pantsar, 2015). Although some remain sceptical about the application of conceptual metaphor theory to describe and explain cognitive processes (e.g., McGlone, 2011), those who support its application typically see it as more than just useful heuristics. On their view, conceptual metaphors are seen as being *essential* for fields like cognitive psychology. As a representative example, consider the way in which Fernandez-Duque and Johnson describe metaphors for attention as "defin[ing] entire research programs, circumscribing which attentional phenomena are studied, how they are studied, and what counts as evidence." (Fernandez-Duque & Johnson, 1999, p. 83). Importantly, the authors identify the rationale of attention metaphors as the cause of problems for candidate theories of attention (Ibid.).

[3] Their account has since been further developed by others, most notably Kövecses (2005, 2010) and Fauconnier and Turner (2002), who developed it into the account of conceptual *blending*. While we do not want to undermine the significance of these later developments, for present purposes the original account of Lakoff and Johnson (2003) is sufficient.

In what follows, we argue that conceptual metaphor theory can help us make sense of the metaphors that are frequently used to describe chatbot mistakes. To be more precise, we argue that it can help us understand why several metaphors used in the literature *fail* to be informative. Just like the metaphors for attention can cause theories of attention to run into problems, as shown by Fernandez-Duque and Johnson (1999), metaphors for chatbot mistakes, we argue, can cause misunderstandings of chatbot behaviours.

But before we can move on to that argumentation, we need to elaborate on conceptual metaphor theory by identifying three ways in which metaphors can fail.[4] The first, and perhaps most obvious, way of failing is when the target domain does not bear sufficient similarity to the source domain. "Snakes are just big worms" is a failed metaphor in this sense. The source domain (worms) is not similar enough to the target domain (snakes) because it does not include important aspects, like being potentially lethal when encountered.

The second way in which conceptual metaphors can fail is when there is no known feasible mapping between the two domains. Here it is important to remember that the conditions for a mapping in conceptual metaphor theory can be – mathematically speaking – rather relaxed. The mapping does not need to be *onto* (so that every item in the target domain is mapped by some item of the source domain), nor does it need to be *injective* (so that every item of the source domain is mapped to a distinct item of the target domain). Indeed, the mapping does not even need to be a *function*, meaning that every item in the source domain is mapped to exactly one element of the target domain. However, for the metaphor to work, some mapping between the two domains is necessary. If it is not feasible to establish a mapping, the metaphor fails. "Mathematics is poetry" can be seen as an example of this. It is possible that the source domain (poetry) and the target domain (mathematics) have sufficient similarity for there to be a mapping from the former to the latter. However, to the best of our knowledge, no good grounds for this have been presented and parallels drawn between mathematics and poetry in the literature are often misguided and confusing.[5]

Finally, a third way in which a metaphor can fail occurs when the source domain itself is not sufficiently clear and unambiguous. Consider, for example, the metaphor "life is hell". The notion of 'hell' is present only in some religions and folklore, and even in these contexts it receives

[4] Curiously, even though a lot has been written about conceptual metaphor theory, we are not aware of an explicit treatment of the different ways in which metaphors can fail.

[5] This is certainly the case with one of the most famous such metaphorical connections, presented by Kristeva (1969). The confused connections between "poetic logic" and mathematics have been highlighted in Bricmont and Sokal (2003).

radically different characterisations. It is thus not clear what, aside from a general but vague impression of extreme negativity, the metaphor is meant to capture. Moreover, many people are likely to have a better understanding of life than of hell, putting the usefulness of the metaphor further in doubt.

On our account, the items on our list of conceptual metaphor failures are not mutually exclusive. Indeed, as we will see in the next section, in many cases conceptual metaphors fail for more than one reason. In all cases, applying the conceptual metaphor in question is not *fruitful*. For one or more reasons, the metaphor does not add to or specify – and is therefore not conducive to – our understanding of the target domain. It should also be noted that the above list of three kinds of conceptual metaphor failures is not meant to be exhaustive. It is theoretically conceivable that conceptual metaphors can fail in other ways. However, we believe that the above account captures the phenomenon of conceptual metaphor failures sufficiently to assess the metaphors used in discussions of the possibilities and limitations of chatbots like ChatGPT.

ChatGPT is typically referred to as an AI application and 'artificial intelligence' is in fact a particularly interesting concept for present purposes. This is because it exemplifies a wider phenomenon in the use of terminology, what Floridi and Nobre (2024) call *conceptual borrowing*. In the development of new disciplines, they argue, terminology is often not invented from scratch, but rather "borrowed" from other disciplines. They list several instances in which there is conceptual borrowing between AI research and the brain and cognitive sciences (Floridi & Nobre, 2024, pp. 5–6). For example, AI research has borrowed the concepts 'embodiment' and 'neuron' from the brain and cognitive sciences, whereas the latter have borrowed the concepts 'circuit' and 'coding' from AI. Floridi and Nobre (2024) argue that this "crosswiring" of the languages between the two disciplines leads to problems, in the worst case to a "conceptual mess" (Floridi & Nobre, 2024, p. 8). This danger, we will argue next, is clearly present in scholarly and popular discourses about chatbot mistakes. The metaphors used to refer to them are cases of conceptual borrowing, and conceptual metaphor theory can help us understand what goes wrong with the borrowed concepts.

5. Mental Metaphors of Chatbot Mistakes

Against the background of conceptual metaphor theory and our identification of three ways in which metaphors can fail, we now proceed to identify, examine, and critically discuss the hallucination, confabulation, bullshitting metaphors that have been employed to capture chatbot mistakes. The overall claim of this section will be that all three mental metaphors fail to be conducive to our understanding of factually incorrect chatbot outputs.

5.1. The Hallucination Metaphor

Recently, scholarly publications across AI research (Lin et al., 2022; Maynez et al., 2020; Zhang et al., 2023), philosophy (Heersmink et al., 2024), medicine (Alkaissi & McFarlane, 2023), and other disciplines (Au Yeung et al., 2023), alongside reports by tech companies (IBM, 2024; OpenAI et al., 2024) and popular media articles (Leffer, 2024; Metz, 2023; Weise & Metz, 2023) have employed the notion of 'hallucination' to refer to the propensity of ChatGPT and other chatbots to output factually mistaken statements (for a review, see Maleki et al., 2024). As a recent example from the growing philosophical literature on LLMs and chatbots, consider the following assumption presented by Heersmink et al. (2024): "Whilst ChatGPT is generally good at generating accurate answers, it does sometimes *hallucinate*, which is impossible to detect if one isn't already knowledgeable on the topic" (p. 10; emphasis added). Furthermore, OpenAI itself notes in its *GPT-4 Technical Report* that "GPT-4 has the tendency to 'hallucinate,' i.e. 'produce content that is nonsensical or untruthful in relation to certain sources'" (OpenAI et al., 2024, p. 46), referencing publications by Maynez et al. (2020) and Lin et al. (2022).[6] Chatbot mistakes are also frequently discussed in the same terminology in articles published in prestigious newspapers and on their online platforms. For example, Weise and Metz (2023) published an article in the *New York Times* titled *When A.I. Chatbots Hallucinate*. In a follow-up piece titled *Chatbots May 'Hallucinate' More Often Than Many Realize*, Metz (2023) notes: "Experts call this chatbot behavior 'hallucination.' [...] Because these chatbots can respond to almost any request in an unlimited number of ways, there is no way of definitively determining how often they hallucinate."

As these examples indicate, the notion of 'hallucination' is frequently used to refer to chatbot mistakes. However, we will argue that this notion is a failed metaphor that impedes a better understanding of the target phenomenon. Hence we go beyond the criticism against the hallucination metaphor recently shared by Østergaard and Nielbo (2023). While they state that

[6] It should be mentioned, however, that the notion of 'hallucination' receives a brief critical appraisal in a footnote: "We use the term 'hallucinations,' though we recognize ways this framing may suggest anthropomorphization, which in turn can lead to harms or incorrect mental models of how the model learns" (OpenAI et al., 2024, p. 46). Yet, the authors of the report continue referring to ChatGPT mistakes as 'hallucinations.'

hallucination is both an “imprecise metaphor” and a “highly stigmatizing metaphor” (Østergaard & Nielbo, 2023, p. 1105), we will show that the metaphor fails entirely, above and beyond its lack of precision and discriminatory potential. To this end, we first outline how the source domain of the metaphor can be delineated in philosophy of mind and the empirical cognitive sciences.

To a first approximation, ‘hallucinations’ can be defined as “perceptual experiences that lack a sensory stimulus” (Wilkinson et al., 2022, p. 219). Similarly, Corlett et al. (2019) define ‘hallucinations’ as “percepts without corresponding external stimuli” (p. 114). Thus defined, hallucinations, understood as a certain kind of perceptual experience, can occur in all sensory modalities (Macpherson, 2013), but visual and auditory hallucinations are most frequently identified (Corlett et al., 2019). Hallucinations can be a positive symptom of a diagnosed or diagnosable mental disorder such as schizophrenia, or a neurodegenerative disease such as Parkinson’s or Alzheimer’s disease (Corlett et al., 2019). Yet, hallucinations are assumed to be frequently experienced also by members of the statistically ‘normal’ population, for example in cases of bereavement (Millar, 2023).

Beyond the initial working definitions of ‘hallucination’ cited above, there is little consensus about the characteristics and scope of the target phenomenon – or range of target phenomena – and their aetiologies within and across disciplines. To begin with, accounts and categorisations of hallucinations differ across cultures and historical times. Already in the contemporary Anglophone context, the specification of the target phenomenon differs across scientific discourses (Wilkinson et al., 2022). In recent philosophy of perception, research has focussed on visual hallucinations (Macpherson, 2013). In philosophy of mind and the empirical cognitive sciences, auditory-verbal hallucinations (*hearing voices*), which are identified as a positive symptom of schizophrenia (Fletcher & Frith, 2009), have received by far the most attention (Corlett et al., 2019; Wilkinson et al., 2022).

In sum, beyond basic working definitions of ‘hallucination’ that can be identified as a common ground within and across disciplines, there are various ways in which this notion can be defined and in which the relevant target phenomenon – or range of target phenomena – can be described and explained. Call this the *conceptual variability problem*. Due to this problem, the hallucination metaphor fails because the source domain is not sufficiently clear and unambiguous.

However, the hallucination metaphor also fails for at least two other reasons. First, there is a common ground in interdisciplinary research that hallucinations qualify as a certain kind of conscious perceptual experiences. By this, we refer to conscious experiences in one or more sensory domains that either have a qualitative character, or are available for cognitive processing, or a combination of both.[7] By definition, then, only systems that can have conscious perceptual experiences can hallucinate. At the time of writing, there is no evidence that ChatGPT or any other chatbot can have perceptual experiences (see Bergstrom & Ogbundu, 2023). Accordingly, the relevant kind of systems in the source domain (human organisms) is categorically different from the relevant kind of systems in the target domain (chatbots) with regards to the occurrence of conscious perceptual experiences, be they hallucinatory or veridical. Call this the *systemic difference problem*. Given this problem, the hallucination metaphor fails because there is no sufficient similarity between the source domain and the target domain.

Second, the hallucination metaphor does not allow for a feasible mapping between the source domain and the target domain. Even if we disregarded the systemic difference problem, the notion of 'hallucination' does not pick out the right kind of states and processes that could be mapped to the target domain. By definition, hallucinations occur during sensory perception. However, chatbot outputs, whether mistaken or factually correct, should be categorised as outputs of language production. In the cognitive sciences, language production, whether in speech or writing, is commonly understood as a particular kind of action (Pickering & Garrod, 2013). Accordingly, hallucinations are not the kind of states that can be feasibly mapped to the target domain, which is defined by active language production, not sensory perception. Call this the *perception-action mismatch problem*.

Taken together, the conceptual variability problem, the systemic difference problem, and the perception-action mismatch problem clearly demonstrate that the hallucination metaphor fails to elucidate chatbot mistakes. The metaphor does not add to our understanding of the target domain or even worse, it can confuse our previous, better, understanding of it.

5.2. The Confabulation Metaphor

We are by no means the first authors to criticise the hallucination metaphor. For example, in the popular online technology outlet *Ars Technica*, Edwards (2023) criticises the hallucination

[7] We are referring here to the now-seminal distinction in philosophy of mind, introduced by Block (1995), between phenomenal consciousness (*p-consciousness*) and access consciousness (*a-consciousness*). 'P-consciousness' refers to the phenomenon of consciously experiencing a state S or process P that has a qualitative character, a "what-it-is-likeness" to use Nagel's (1974) terminology. The notion of 'a-consciousness' captures the phenomenon of consciously experiencing a state S or a process P that might lack a qualitive character, but is available for further cognitive processing (e.g., rational deliberation).

metaphor for being inadequate. In its stead, referencing a neurological review article of research on confabulation by Brown et al. (2017), he proposes using ‘confabulation’ as a metaphor to capture chatbot mistakes (for a critical discussion, see Hicks et al., 2024):

> [...] we feel the term ‘confabulation,’ although similarly imperfect, is a better metaphor than ‘hallucination.’ 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 [...]. [...] Key to understanding ChatGPT's confabulation ability is understanding its role as a prediction machine. When ChatGPT confabulates, it is reaching for information or analysis that is not present in its data set and filling in the blanks with plausible-sounding words. (Edwards, 2023)

Similarly, on the popular science website *Psychology Today*, the psychologist Henriques (2024) proposes to replace the hallucination metaphor with the confabulation metaphor to capture chatbot mistakes:

> If hallucinations are not the right word for these errors, what is a better word? If we are pulling from the world of psychiatry and clinical psychology, a much better word is confabulation. Confabulation is when individuals generate false content without meaning to deceive, often to fill in some expected social role. [...] What [chatbots] are doing is generating language modeling [sic!] that is nonsensical or false, without any intent to deceive, based on problems with retrieval, source, and comprehension.

Furthermore, the confabulation metaphor features in the discourse of healthcare professionals who explore the possibilities and limitations of using ChatGPT for the education and training of medical practitioners (Rodgers et al., 2023). While less prevalent than the hallucination metaphor, the confabulation metaphor is still used frequently enough to warrant a careful assessment of its appropriateness and epistemic usefulness.

The state of research on confabulation resembles in many ways the landscape of hallucination research. The notion of ‘confabulation’, its conceptual scope, and its contributions to psychiatric discourse formation have varied considerably within and across Western cultures since the beginning of its frequent use at the turn from the 19th to the 20th century (Berrios, 1998). In contemporary philosophy of mind and the cognitive sciences, the notion of ‘confabulation’ remains polysemic. Hirstein (2006), for example, proposes to distinguish between a *mnemonic* *concept*, a *linguistic concept*, and an *epistemic concept*, each highlighting certain aspects of behaviour that can be identified in some, but not all cases of confabulation.

Some basic working definitions of 'confabulation' have been developed that try to integrate relevant aspects of the phenomenon across all three conceptual dimensions. However, these working definitions are far from uncontroversial (Hirstein, 2009). As an example, consider the following considerations by Bortolotti and Cox (2009):

> Most typically, people confabulate when they make statements or tell stories which might be either inaccurate or badly supported by the available evidence. The 'story' is genuinely believed by the subject reporting it; it can also be endorsed with some conviction, and maintained in the face of counterarguments. This characterisation of confabulation has generated controversy because it is too vague unless it can be further qualified according to one's preferred hypothesis about how confabulation arises. (Bortolotti & Cox, 2009, p. 952)

15 years after Bortolotti and Cox (2009) identified this *definition problem*, as they call it, it still remains unclear how 'confabulation' can be defined within and across contexts. Confabulation has been identified as a symptom of psychiatric disorders, such as misidentification syndromes (e.g., Capgras syndrome) and anosognosia (e.g., Anton's syndrome) as well as neurodegenerative disorders such as dementia (Berrios, 1998; Hirstein, 2006, 2009). More recently, however, it has been argued that confabulation frequently occurs in members of the statistically 'normal' population. These include young children, participants in experiments tapping into decision-making or moral reasoning, and hypnotised adults that have not received a psychiatric diagnosis (Bortolotti, 2018; Bortolotti & Cox, 2009; Coltheart, 2017; Hirstein, 2006). This casts doubt on the possibility that a general, sufficiently specific and informative understanding of 'confabulation' can be provided (for a critical discussion, see Robins, 2020).

In sum, then, philosophical and empirical research indicates that the notion of 'confabulation' is polysemic: it takes on different meanings depending on its conceptual and descriptive scope, the range of phenomena to which it can be applied, and the epistemic, aetiological, and pragmatic aspects that are taken to be most important for adequately capturing relevant clinical, experimental, and everyday communicative behaviours. In this sense, like the hallucination metaphor, the confabulation metaphor faces the *conceptual variability problem*. Since the source domain of 'confabulation' is not sufficiently clear and unambiguous, the confabulation metaphor fails.

Yet, perhaps there could be a way to save the confabulation metaphor from failure. After all, Edwards (2023) seems to narrow down his understanding of ‘confabulation’ by picking out a mnemonic conceptualisation. Recall that according to Edwards (2023), “[...] a ‘confabulation’ occurs when someone's memory has a gap and the brain convincingly fills in the rest without intending to deceive others.” While acknowledging the inaccurate and overly simplifying character of this statement, we could concede that Edwards (2023) identifies a particular kind of confabulation, namely *mnemonic confabulation*, as the relevant source domain for the confabulation metaphor. This conceptual restriction would avoid the conceptual variability problem. In the philosophy of memory, a mnemonic confabulation is defined as follows: “Mnemonic confabulation occurs when there is no relation between a person’s seeming to remember a particular event or experience and any event or experience from their past – either because there is no such event in their past or because any similarity to such an event is entirely coincidental” (Robins, 2020, pp. 125–126).

As this definition makes clear, mnemonic confabulation occurs when human agents engage in episodic remembering – when they seem to remember a particular event or experience from their own personal past. However, to the best of our current knowledge, ChatGPT and other chatbots do not experience personally relevant events that they could subsequently seem to remember episodically. So even if we restricted the metaphor to mnemonic confabulation, the confabulation metaphor would fail. While the mnemonic confabulation metaphor would avoid the conceptual variability problem, it would fail because it faces the *systemic difference problem*. The relevant kind of systems in the source domain (human organisms) is categorically different from the relevant kind of systems in the target domain (chatbots) with regards to the ability to episodically remember personal past events and experiences. Hence, the source domain does not bear sufficient similarity to the target domain in order for the metaphor to work.[8]

Ultimately, the systemic difference problem makes the confabulation metaphor fail also for another reason. There is an emerging consensus in philosophy and the cognitive sciences that a key purpose of confabulation consists in a (largely unconscious) attempt to engage in communicative acts that are conducive to meaning-making and an understanding of self-relevant events and experiences (Bortolotti, 2018; Bortolotti & Cox, 2009; Coltheart, 2017; Stammers, 2020). This purpose is deeply rooted in human practices of social interaction and

[8] While this ability is not acquired in all cases throughout ontogenetic development, and can be (partially) lost due to trauma or neurodegenerative diseases, it is an ability that is acquired by most humans and maintained throughout their lives in the vast majority of cases.

communication. At the time of writing, there is no evidence available to support the assumption that ChatGPT, or any other chatbot, has a propensity to engage in communicative activities that are conducive to meaning-making and practices of understanding of self-significant events or experiences. Whatever the processes are that give rise to chatbot mistakes, they are not similar to the kind of processes that human agents engage in when they confabulate in the context of social interactions and conversations with other human agents. For this reason, the relevant kind of process in the source domain (a confabulatory, communicative act of a human agent) is categorically different from the relevant kind of process in the target domain (a chatbot's generation of a factually incorrect output). Accordingly, the confabulation metaphor fails because the source domain is not sufficiently similar to the target domain.

To sum up, because of the conceptual variability problem, and two variants of the systemic difference problem, the confabulation metaphor fails. Consequently, it obscures, rather than adds to or specifies our understanding of the target domain.

5.3. The Bullshitting Metaphor

Refuting the notions of 'hallucination' and 'confabulation' to metaphorically capture ChatGPT mistakes, Bergstrom and Ogbundu (2023) and Hicks et al. (2024) propose to conceive of these mistakes as 'bullshit' in Frankfurt's (2005) sense. In what follows, we will focus on Hicks's et al. (2024) account, because this is currently the most developed version of the bullshitting metaphor. In his now-seminal work, Frankfurt (2005) offers a conceptual analysis of the vernacular notions of 'bullshit', 'bullshitting', and 'bullshitter' which is contrasted with 'lie', 'lying', and 'liar', respectively. For present purposes, the difference between the 'bullshitter' and the 'liar' is key to understanding the proposal of Hicks et al. (2024):

> This is the crux of the distinction between [the bullshitter] and the liar. Both he and the liar represent themselves falsely as endeavouring to communicate the truth. The success of each depends upon deceiving us about that. But the fact about himself that the liar hides is that he is attempting to lead us away from a correct apprehension of reality; we are not to know that he wants us to believe something he supposes to be false. The fact about himself that the bullshitter hides, on the other hand, is that the truth-values of his statements are of no central interest to him; what we are not to understand is that his intention is neither to report the truth nor to conceal it. This does not mean that his speech is anarchically impulsive, but that the motive guiding and controlling it is unconcerned with how the things about which he speaks truly are. (Frankfurt, 2005, pp. 54–55)

In other words, both the liar and the bullshitter are in the business of communicating statements that do not truthfully represent states of affairs in the real world. However, while the liar is concerned with the truth, only to be able to conceal it, the bullshitter's utterances display an "indifference to how things really are" (Ibid., p. 34). Amongst the general notion of 'bullshit', Hicks et al. (2024) distinguish between two species. *Hard bullshit* is produced with the communicative intention to mislead the audience about the utterer's agenda. By contrast, *soft bullshit* is produced without the communicative intention to mislead the hearer regarding the utterer's agenda (Hicks et al., 2024, p. 5).

Hicks et al. (2024) argue that, at the very least, ChatGPT frequently generates soft bullshit: "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*" (Hicks et al., 2024, p. 6). However, there are at least two ways in which the soft bullshitting metaphor induces the *systemic difference problem*. First, this metaphor comes at the cost of anthropomorphising ChatGPT, for Hicks et al. (2024) explicitly state that it is "engaged in the business of outputting utterances" that do not reveal its indifference towards their truthfulness. But ChatGPT and other chatbots are not engaged in *anything* in a sense that is similar to human engagements in activities. Second, the definition of 'soft bullshit' includes the assumption that the communicating system under consideration has an "agenda". However, just as ChatGPT does not engage in anything, it also does not have an agenda. For these two reasons, we assume that the soft bullshitting metaphor is yet another instantiation of the *systemic difference problem*, because communicative outputs of human agents (the source domain) and ChatGPT (the target domain) substantially differ regarding their contextualisation in engaged activities and agendas. Accordingly, this metaphor fails because there is no sufficient similarity between the source and target domains.

However, Hicks et al. (2024) go one step further and discuss under what circumstances ChatGPT could be considered as a *hard bullshitter*. They assume that this consideration is apt if we apply the intentional stance in a Dennettian (1987) sense.[9] While Hicks et al. (2024) concede that

[9] Dennett (1987) proposed to distinguish between three relevant stances to predict or explain the behaviour of biological and non-biological systems. First, in the *physical stance*, one tries to predict or explain a system's behaviour by collecting and integrating information about its physical properties and the laws of nature that govern them. Second, in the *design stance*, one tries to predict or explain a system's behaviour by referring to its design principles and functional configuration. Finally, in the *intentional stance*, one tries to predict or explain a system's behaviour by ascribing beliefs and desires, and by establishing rational connections between those propositional states and the system's behaviour. Importantly, according to Dennett (1987), the intentional stance works irrespective of whether or not the target system actually enjoys propositional states such as beliefs and desires.

ChatGPT does not have intentions in any proper sense of the word, they assume that it is configured in such a way that human agents can apply intentional notions to describe its outputting behaviour: "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" (Hicks et al., 2024, p. 8; for a similar point, see Bergstrom & Ogbundu, 2023).

However, this assumption is problematic for two reasons. First, it changes the topic from *having* intentions to *ascribing* intentions. Recall that hard bullshit is defined as bullshit that is "produced with the intention to mislead the audience about the utterer's agenda" (Ibid., p. 5). According to this definition, hard bullshitting occurs if a certain statement is intentionally uttered, irrespective of whether or not a recipient describes the relevant communicative behaviour as intentional. However, the possibility to describe ChatGPT's outputting behaviour in intentional terms does not entail that it misleads its human interlocutors intentionally and therefore qualifies as a hard bullshitter.

Second, the assumption of Hicks et al. (2024) changes the target domain from ChatGPT's outputting behaviour to the intentions of its developers. The view that the developers at OpenAI intentionally devised ChatGPT such that its (main) task consists in deceiving human users about its communicative purpose remains speculative at best.[10] At the time of writing, there is no evidence of the purposes and intentions of ChatGPT's developers that could corroborate such an assumption. But even if such evidence existed, the explanandum consists in ChatGPT mistakes, not the attitudes of the human agents who developed the chatbot.

In sum, while Hicks et al. (2024) set out to improve scholarly and popular discussions of ChatGPT mistakes by replacing the hallucination and confabulation metaphors with the bullshitting metaphor, their positive proposal does not succeed. If 'soft bullshitting' is selected as the source domain, the metaphor will fail because there is no sufficient similarity between the source domain and the target domain due to the systemic difference problem. If 'hard bullshitting' is chosen as the source domain, the metaphor will fail because Hicks et al. (2024) change the topic from ChatGPT (the target domain) *having* intentions to behaving such that intentions can be

[10] Here Hicks et al. (2024) seem to interpret ChatGPT in the context of the Turing Test. In the original version of the Turing Test, machines are tested on whether human interrogators can distinguish their output from that of a human interlocutor (Turing, 1950). While it is clear that ChatGPT outputs are designed to be human-like, there is no evidence to suggest that it is designed to deceive users about its communicative purposes.

ascribed. It will also fail because they change the target domain from ChatGPT's outputting behaviour to the intentions of its designers. Ultimately, both versions of the bullshitting metaphor as proposed by Hicks et al. (2024) fail because they do not add to or specify our understanding of the target domain.

6. Discussion and Concluding Remarks

In the previous section, we argued that the mental metaphors, which are used by researchers, AI developers, and journalists to discuss factual mistakes in chatbot outputs, fail to be conducive to a better understanding of the target phenomenon. Our analysis also reveals a pattern. The use of the hallucination metaphor received criticism and was replaced by the confabulation metaphor (Edwards, 2023; Henriques, 2024). In turn, the confabulation metaphor was replaced by the bullshitting metaphor (Hicks et al., 2024). In this paper, we want to break free from that pattern. Rather than continuing the iterative process of finding an arguably better metaphor, we want to question the very process of capturing and discussing chatbot mistakes by using metaphors drawn from human perception, action, and cognition. But why have such metaphors been so popular within and across different types of scholarly and popular discourse? In this section, we identify five reasons for using mental metaphors. By way of responding to two possible objections, we will then argue that the use of these metaphors is likely to continue creating epistemic confusions about chatbot mistakes.

The first reason for the popularity of mental metaphors has been well known in AI research since the 1960s, and it is the tendency to *anthropomorphise* machines. Already in the case of ELIZA (see Section 2), its developer noted that it was difficult to convince some people that the computer programme was not human (Weizenbaum, 1966, p. 42). When human subjects are given the task of identifying humans and computers based on textual outputs, it is more common to mis-identify the computers as humans than vice versa (Copeland, 2000, p. 525). Given such tendencies, it is not surprising that notions drawn from descriptions of the human mind are used to capture the functioning of computers. As noted in Section 4, this kind of conceptual borrowing is commonplace when terminology is needed in (relatively) new disciplines (Floridi & Nobre, 2024). Given that chatbots like ChatGPT are designed to provide human-like outputs, it is to be expected that phenomena like factual output mistakes are described in terminology drawing from discourses about the human mind. All three metaphors discussed in this paper seem to involve this kind of anthropomorphisation.

Second, an important reason for using mental metaphors concerns Dennett's (1987) notion of the *intentional stance* (see Section 5.3). Ascribing intentions, beliefs and desires can be beneficial for explaining the behaviour of an artificial system. However, it is important to remember that adopting the intentional stance comes with no guarantee of improved understanding. While it is possible to adopt the intentional stance for explaining mistakes in chatbot outputs, it can lead to conceptual confusions. There is no general, *prima facie* advantage of adopting the intentional stance across the board and its epistemic usefulness varies considerably across cases. Our analysis of the bullshitting metaphor suggests that the intentional stance is inherently problematic for understanding or explaining factual chatbot mistakes.

The third reason for using mental metaphors, we submit, comes from the success that the use of metaphors, in general, often has in capturing and explaining new phenomena. Conceptual metaphors, as described by Lakoff and Johnson (2003), are often highly beneficial for understanding new domains. Similarly, conceptual borrowing as specified in Floridi and Nobre (2024) can lead to improved understanding of new phenomena. The use of conceptual metaphors is a valid tool for scientific discourse, and we definitely do not want to suggest otherwise. However, each use of conceptual metaphors is its own case (for an example, see Fernandez-Duque & Johnson, 1999). It does not come with guaranteed epistemic power.[11]

Fourth, as we have seen, metaphorical language is sometimes encouraged by the developers of ChatGPT and other chatbots. This is also likely to be an important reason for the popularity of engaging with metaphors in scientific and philosophical discourses. As mentioned in Section 5.1, OpenAI itself regularly uses the concept of 'hallucination' in discussing ChatGPT's functionality. They recognise this kind of talk as being a case of anthropomorphisation, yet seem to have no intention of letting it go. We suspect that an economic incentive may be at play here. Chatbot mistakes are inevitable with the current architecture, but tech companies are perhaps not eager to acknowledge them in those terms. Hence 'hallucination' may be kind of a sleight of hand from the developers, used to hide a weakness of their product on a highly competitive market of AI applications.

Finally, the use of mental metaphors to capture chatbot mistakes is likely to arise partly from the way chatbot outputs are discussed in general. Human agents are typically enculturated to

[11] This does not mean, however, that we cannot make inductive generalisations about applying a certain type of source domain to a particular target domain. Indeed, the failures of the three mental metaphors identified and discussed in this paper can be seen as an inductive argument against using the source domain of the human mind to characterise mistaken chatbot outputs metaphorically.

ascribe properties like truth-orientedness, trust, intention, and reliability to their interlocutors. Unless there is a good reason to think otherwise, ascribing such properties may well be the default position when communicating with other agents. Hence, whether ChatGPT is correct or not in its outputs, it is to be expected that the user gets the impression that the chatbot is interested in or committed to making true statements.[12]

While the above reasons help explain why mental metaphors are widely used to describe factual mistakes in chatbot outputs, none of them should be seen as a *justification* for this habit. At this point, however, one might ask whether we over-emphasise or exaggerate the issue of metaphoric language. After all, one might argue, it is simply a question of language use. If the metaphoric discourse illuminates the issue of chatbot mistakes for at least some people, where is the harm? There are four reasons why this *mere language use objection* fails. First, the implicit consensus in the relevant literatures seems to be that metaphors are not simply components of linguistic practices without further importance. This is apparent in the way authors point out problems when it comes to *other*, competing metaphors. As we have seen, proponents of the confabulation metaphor tend to see the hallucination metaphor as problematic, and proponents of the bullshitting metaphor reject both the confabulation and the hallucination metaphor. This suggests that the problematic recruitment of metaphors is generally not seen as acceptable – correctly, in our view.

The second reason why the *mere language use objection* fails is that our terminology, including the metaphors we use, is important for understanding and explaining scientific target phenomena. The key tenet of conceptual metaphor theory is that metaphors are not a merely linguistic phenomenon. Instead, they are a key part of the cognitive process of gaining epistemic access to new domains. Using failed metaphors can damage or impede this process, for they can obscure relevant aspects of the explanandum.

Related to this is the third reason why the *mere language use objection* fails. There is a risk that not everybody understands statements such as "ChatGPT hallucinates", "ChatGPT confabulates" or "ChatGPT is bullshitting" *as being* metaphorical. Adopting the intentional stance, it is commonplace to speak of the intentions and desires of AI systems. There is of course nothing new in that: Dennett introduced the notion of the 'intentional stance' in 1987 when the state of technology was quite different. However, in this difference lies a danger. Whereas most

[12] This impression that the chatbot is somehow invested in generating true statements is, of course, strengthened by the outputted language use. For example, ChatGPT often generates "apologies" in its next output when the user points out that it was wrong.

people are likely to understand that, say, a thermostat does not really have intentions or desires, the matter may not be so clear with contemporary AI applications. Given the human tendency to anthropomorphise machines, there is a danger that users start genuinely believing that an AI system has perceptions, intentions, beliefs, desires, and perhaps even feelings (Christoforakos et al., 2021). This can become problematic if chatbot outputs are treated like those of human conversational partners, for example, as being able to offer well-informed and socio-culturally informed advice on emotionally sensitive issues. These concerns have already been raised by researchers working on social chatbots (e.g., Brandtzaeg et al., 2022; Skjuve et al., 2022; Weber-Guskar, 2022) and therapy chatbots (e.g., Tekin, 2021). In such cases, it would be important for the users to recognise and be reminded that chatbots do not possess emotions or the ability to feel empathy. While ascribing emotions may be a more serious problem, we submit that already ascribing intentions, or genuine forms of understanding (Mitchell & Krakauer, 2023), to the chatbot can mischaracterise human-chatbot exchanges in potentially damaging ways.

Finally, the ascription of human characteristics to ChatGPT and other chatbots can also exacerbate the tendency to gradually lose the ability to distinguish between human agents and artificial systems. Using mental metaphors may be conducive to this kind of development. If we come to think that chatbots and other artificial systems can, indeed, hallucinate, confabulate, or bullshit just like humans, there is a risk that we lose sight of the genuine abilities and limitations of both human *and* artificial systems. This is another reason why we should take the metaphors we use very seriously.

A critic of our position might concede that we succeed in rejecting the *mere language use objection* to a sufficient degree. Yet, they might object to our strategy to capture the source domain (the human mind) in philosophical and scientific terms. Instead, they might argue, we should be operating with a simpler, folk-psychological understanding of it. According to this potential objection, we impose too strict constraints on the clarity and unambiguity of the source domain when the metaphors are actually based on folk-psychological concepts. Call this the *folk-psychological concepts objection*. Our reply to this objection is twofold. First, the metaphors at issue, which capture factual chatbot mistakes as hallucinations, confabulations, or bullshit, are frequently used in philosophical and scientific discourses on ChatGPT and other chatbots. The question then is why research in philosophy, AI, and cognate disciplines should be informed or even guided by metaphors that pick out a folk-psychologically conceived source domain. *Prima facie*, philosophical and scientific concepts are more specific, precise, and well-developed than their folk-psychological counterparts. It stands to reason that research in relevant disciplines should aim for conceptual clarity and avoid conceptual ambiguity.

Consequently, operating with folk-psychological concepts is difficult to motivate in scholarly contexts when more precise, theoretically formed and scientifically corroborated concepts are available.

Second, even where the metaphors at issue are used outside philosophical-scientific discourses and are disseminated through newspaper articles, popular science pieces, or progress reports of big tech companies, the *folk-psychological concepts objection* does not succeed. As research in experimental philosophy and philosophy of language shows, human agents have different understandings of key folk-psychological concepts such as ‘vision’ (Fischer et al., 2023), ‘pain’ (Coninx et al., 2024; Liu, 2023), or ‘lying’ (Wiegmann & Meibauer, 2019). This makes it unlikely that *the folk* has a shared and ubiquitous understanding of other folk-psychological concepts such as ‘hallucination’, ‘confabulation’, and ‘bullshitting.’ This lack of a folk-psychological consensus about the relevant source domains is exacerbated by empirical evidence suggesting that folk-psychological concepts differ substantially across cultures (Henrich et al., 2010). Accordingly, the hallucination, confabulation, and bullshitting metaphors fail also when their source domains are situated in the context of folk-psychology. The conceptual variability problem still applies, as folk-psychological concepts tend to be as polysemous as their philosophical-scientific counterparts.

Moving forward, what kind of language should we use to discuss factual mistakes in chatbot outputs? While we do not object to the use of metaphors in principle, we want to challenge a (largely implicit) background assumption in much recent research on ChatGPT and other chatbots. According to this assumption, metaphors are needed to understand chatbot mistakes. However, we invite philosophers and AI researchers to explain and justify *why* we should keep using a particular metaphor – or indeed why we should identify and apply metaphors in the first place. What exactly is gained by using metaphorical language that cannot be achieved by characterising chatbot outputs with descriptive – albeit unexciting – terms like ’undesired’ ’factually incorrect’, ’inaccurate’, or ’wrong’? What is gained by talking about a chatbot as if it were an intentional agent? Do we run risk of missing out on important insights if we discuss ChatGPT simply as a computer programme? As we have argued in this paper, our position in response to these last two questions is clear: the hallucination, confabulation, and bullshitting metaphors obscure, rather than enable a proper, epistemically fruitful understanding of ChatGPT and other chatbots, as well as the limitations of AI systems more generally. What we would gain from eliminating these mental metaphors is an increased conceptual and scientific understanding of chatbot mistakes – and their implications for practices of knowledge generation and meaning-making.

References

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

Au Yeung, J., Kraljevic, Z., Luintel, A., Balston, A., Idowu, E., Dobson, R. J., & Teo, J. T. (2023). AI chatbots not yet ready for clinical use. *Frontiers in Digital Health*, *5*. https://doi.org/10.3389/fdgth.2023.1161098

Bassett, C. (2019). The computational therapeutic: Exploring Weizenbaum's ELIZA as a history of the present. *AI & SOCIETY*, *34*(4), 803–812. https://doi.org/10.1007/s00146-018-0825-9

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

Bergstrom, C. T., & Ogbundu, C. B. (2023). ChatGPT isn't "hallucinating." It's bullshitting. *Undark*. https://undark.org/2023/04/06/chatgpt-isnt-hallucinating-its-bullshitting/

Berrios, G. E. (1998). Confabulations: A conceptual history. *Journal of the History of the Neurosciences*, *7*(3), 225–241.

Biswas, S. (2023). Evaluating Errors and Improving Performance of ChatGPT. *International Journal of Clinical and Medical Education Research*, *2*(6), 182–188.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188

Bortolotti, L. (2018). Stranger than fiction: Costs and benefits of everyday confabulation. *Review of Philosophy and Psychology*, *9*(2), 227–249. https://doi.org/10.1007/s13164-017-0367-y

Bortolotti, L., & Cox, R. E. (2009). 'Faultless' ignorance: Strengths and limitations of epistemic definitions of confabulation. *Consciousness and Cognition*, *18*(4), 952–965. https://doi.org/doi:10.1016/j.concog.2009.08.011

Bose, S. (2023). *Bug vs Error: Key Differences*. BrowserStack. https://browserstack.wpengine.com/guide/difference-between-bugs-and-errors/

Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research*, *48*(3), 404–429. https://doi.org/10.1093/hcr/hqac008

Bricmont, J., & Sokal, A. (2003). *Intellectual Impostures* (Main-Re-issue edition). Profile Books.

Brown, J., Huntley, D., Morgan, S., Dodson, K., & Cich, J. (2017). Confabulation: A guide for mental health professionals. *International Journal of Neurology and Neurotherapy*, *4*(2), 4:070.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Christoforakos, L., Feicht, N., Hinkofer, S., Löscher, A., Schlegl, S. F., & Diefenbach, S. (2021). Connect with me. Exploring influencing factors in a human-technology relationship based on regular chatbot use. *Frontiers in Digital Health*, *3*. https://www.frontiersin.org/articles/10.3389/fdgth.2021.689999

Coltheart, M. (2017). Confabulation and conversation. *Cortex*, *87*, 62–68. http://dx.doi.org/10.1016/j.cortex.2016.08.002

Coninx, S., Willemsen, P., & Reuter, K. (2024). Pain linguistics: A case for pluralism. *The Philosophical Quarterly*, *74*(1), 145–168. https://doi.org/10.1093/pq/pqad048

Copeland, B. J. (2000). The Turing Test*. *Minds and Machines*, *10*(4), 519–539. https://doi.org/10.1023/A:1011285919106

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers, A. R., III. (2019). Hallucinations and strong priors. *Trends in Cognitive Sciences*, *23*(2), 114–127. https://doi.org/10.1016/j.tics.2018.12.001

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

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in chatgpt: Analyzing persona-assigned language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1236–1270). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.88

Edwards, B. (2023). Why ChatGPT and Bing Chat are so good at making things up. *Ars Technica*. https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/

Fauconnier, G., & Turner, M. (2002). *The Way We Think: Conceptual Blending And The Mind's Hidden Complexities*. Basic Books.

Fernandez-Duque, D., & Johnson, M. L. (1999). Attention metaphors: How metaphors guide the cognitive psychology of attention. *Cognitive Science*, *23*(1), 83–116.

Fischer, E., Allen, K., & Engelhardt, P. E. (2023). Fragmented and conflicted: Folk beliefs about vision. *Synthese*, *201*(3), 84. https://doi.org/10.1007/s11229-023-04066-w

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48–58. https://doi.org/10.1038/nrn2536

Floridi, L., & Nobre, A. C. (2024). Anthropomorphising Machines and Computerising Minds: The Crosswiring of Languages between Artificial Intelligence and Brain & Cognitive Sciences. *Minds and Machines*, *34*(1), 5. https://doi.org/10.1007/s11023-024-09670-4

Frankfurt, H. G. (2005). *On bullshit*. Princeton University Press.

Gonçalves, B. (2023). The Turing Test is a Thought Experiment. *Minds and Machines*, *33*(1), 1–31. https://doi.org/10.1007/s11023-022-09616-8

Heersmink, R., de Rooij, B., Clavel Vázquez, M. J., & Colombo, M. (2024). A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. *Ethics and Information Technology*, *26*(3), 41. https://doi.org/10.1007/s10676-024-09777-3

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–135. https://doi.org/10.1017/S0140525X0999152X

Henriques, G. (2024). Chatbots do not hallucinate, they confabulate. *Psychology Today*. https://www.psychologytoday.com/us/blog/theory-of-knowledge/202403/chatbots-do-not-hallucinate-they-confabulate

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, *26*(2), 38. https://doi.org/10.1007/s10676-024-09775-5

Hirstein, W. (2006). *Brain fiction: Self-deception and the riddle of confabulation*. MIT Press.

Hirstein, William. (2009). Introduction: What is confabulation? In W. Hirstein (Ed.), *Confabulation: Views from neuroscience, psychiatry, psychology and philosophy* (pp. 1–12). Oxford University Press.

IBM. (2024). *What are AI hallucinations?* https://www.ibm.com/topics/ai-hallucinations

Kasirzadeh, A., & Gabriel, I. (2023). In conversation with artificial intelligence: Aligning Language Models with human values. *Philosophy & Technology*, *36*(2), 27. https://doi.org/10.1007/s13347-023-00606-x

Kövecses, Z. (2005). *Metaphor in Culture: Universality and Variation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511614408

Kövecses, Z. (2010). *Metaphor: A practical introduction* (2nd ed). Oxford University Press.

Kristeva, J. (1969). *Sèméiotikè. Recherches pour une sémanalyse*. Éditions du Seuil. https://www.amazon.de/-/en/Julia-Kristeva/dp/2020019507

Lakoff, G., & Johnson, M. (2003). *Metaphors we live by*. University of Chicago Press.

Lakoff, G., & Núñez, R. (2000). *Where mathematics comes from*. Basic Books.

Leffer, L. (2024). AI chatbots will never stop hallucinating. *Scientific American*. https://www.scientificamerican.com/article/chatbot-hallucinations-inevitable/#:~:text=One%20main%20reason%20AI%20chatbots,of%20Illinois%20at%20Urbana%2DChampaign

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229

Liu, M. (2023). The polysemy view of pain. *Mind & Language*, *38*(1), 198–217. https://doi.org/10.1111/mila.12389

Macpherson, F. (2013). The philosophy and psychology of hallucination: An introduction. In F. Macpherson & D. Platchias (Eds.), *Hallucination: Philosophy and psychology* (pp. 1–38). The MIT Press.

Maleki, N., Padmanabhan, B., & Dutta, K. (2024). *AI hallucinations: A misnomer worth clarifying*. https://arxiv.org/abs/2401.06796

Mauldin, M. L. (1994). Chatterbots, tinymuds, and the turing test entering the loebner prize competition. *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, 16–21.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). *On faithfulness and factuality in abstractive summarization*. https://arxiv.org/abs/2005.00661

McGlone, M. S. (2011). Hyperbole, Homunculi, and Hindsight Bias: An Alternative Evaluation of Conceptual Metaphor Theory. *Discourse Processes*, *48*(8), 563–574. https://doi.org/10.1080/0163853X.2011.606104

Metz, C. (2023). Chatbots may ‘hallucinate’ more often than many realize. *The New York Times*. https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html

Millar, B. (2023). Grief’s impact on sensorimotor expectations: An account of non-veridical bereavement experiences. *Phenomenology and the Cognitive Sciences*, *22*(2), 439–460. https://doi.org/10.1007/s11097-021-09759-6

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI’s large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 Technical Report*. https://arxiv.org/abs/2303.08774

Østergaard, S. D., & Nielbo, K. L. (2023). False responses from Artificial Intelligence models are not hallucinations. *Schizophrenia Bulletin*, *49*(5), 1105–1107. https://doi.org/10.1093/schbul/sbad068

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pantsar, M. (2015). In search of aleph-null: How infinity can be created. *Synthese*, *192*(8), Article 8.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(04), 329–347. https://doi.org/10.1017/S0140525X12001495

Robins, S. (2020). Mnemonic confabulation. *Topoi*, *39*(1), 121–132. https://doi.org/10.1007/s11245-018-9613-x

Rodgers, D. L., Needler, M., Robinson, A., Barnes, R., Brosche, T., Hernandez, J., Poore, J., VandeKoppel, P., & Ahmed, R. (2023). Artificial Intelligence and the simulationists. *Simulation in Healthcare*, *18*(6). https://journals.lww.com/simulationinhealthcare/fulltext/2023/12000/artificial_intelligence_and_the_simulationists.7.aspx

Sheremeta, O. (2023, June 25). Bug vs. Defect: Difference With Definition Examples Within Software Testing. *Testomat.Io*. https://testomat.io/blog/bug-vs-defect-difference-with-definition-examples-within-software-testing/

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2022). A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies*, *168*, 102903. https://doi.org/10.1016/j.ijhcs.2022.102903

Stammers, S. (2020). Confabulation, explanation, and the pursuit of resonant meaning. *Topoi*, *39*(1), 177–187. https://doi.org/10.1007/s11245-018-9616-7

Tekin, Ş. (2021). Is Big Data the new stethoscope? Perils of digital phenotyping to address mental illness. *Philosophy & Technology*, *34*(3), 447–461. https://doi.org/10.1007/s13347-020-00395-7

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *LIX*(236), Article 236. https://doi.org/10.1093/mind/LIX.236.433

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Weber-Guskar, E. (2022). Reflecting (on) Replika. Can we have a good affective relationship with a social chatbot? In J. Loh & W. Loh (Eds.), *Social Robotics and the Good Life* (pp. 103–126). transcript Verlag. https://doi.org/10.1515/9783839462652

Weise, K., & Metz, C. (2023). When A.I. chatbots hallucinate. *The New York Times*. https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine | Communications of the ACM. *Communications of the ACM*, *9*(1), 36–45.

Wiegmann, A., & Meibauer, J. (2019). The folk concept of lying. *Philosophy Compass*, *14*(8), e12620. https://doi.org/10.1111/phc3.12620

Wilkinson, S., Green, H., Hare, S., Houlders, J., Humpston, C., & Alderson-Day, B. (2022). Thinking about hallucinations: Why philosophy matters. *Cognitive Neuropsychiatry*, *27*(2–3), 219–235. https://doi.org/10.1080/13546805.2021.2007067

Zhang, M., Press, O., Merrill, W., Liu, A., & Smith, N. A. (2023). *How language model hallucinations can snowball*. https://arxiv.org/abs/2305.13534