Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

AI Assertion

PATRICK BUTLIN
University of Oxford

EMANUEL VIEBAHN
University of Hamburg

Modern generative AI systems have shown the capacity to produce remarkably fluent language, prompting debates both about their semantic understanding and, less prominently, about whether they can perform speech acts. This paper addresses the latter question, focusing on assertion. We argue that to be capable of assertion, an entity must meet two requirements: it must produce outputs with descriptive functions, and it must be capable of being sanctioned by agents with which it interacts. The second requirement arises from the nature of assertion as a norm-governed social practice. Pre-trained large language models that have not been subject to fine-tuning fail to meet the first requirement. Language models that have been fine-tuned for "groundedness" or "correctness" may meet the first requirement, but fail the second. We also consider the significance of the point that AI systems can be used to generate proxy assertions on behalf of human agents.

Keywords: artificial intelligence; generative AI; large language models; AI pragmatics; assertion; speech acts; sanctions

1. Introduction

The development of Transformer-based large language models (LLMs) and LLM-driven conversational agents has put language use at the cutting edge of AI research. The measurable performance of these systems is human-like on many linguistic tasks (Mahowald & Ivanova et al. 2023), and their capacity to produce fluent and apt outputs is extraordinary. These systems include GPT-3 (Brown et al. 2020) and several successors, including ChatGPT (OpenAI 2022) and GPT-4 (OpenAI 2023). They also include LaMDA (Thoppilan et al. 2022),

Contact: Patrick Butlin <patrick.butlin@googlemail.com>
Emanuel Viebahn <emanuel.viebahn@uni-hamburg.de>

Sparrow (Glaese et al. 2022) and others. In this paper we use LaMDA and Sparrow as our primary examples. The success of these systems raises several pressing questions about the nature of their linguistic and communicative capacities.

Some of these questions concern the *semantic* capabilities of AI systems such as LaMDA. LaMDA famously produced the following output in response to a question about its consciousness/sentience:

> (1) The nature of my consciousness/sentience is that I am aware of my existence, I desire to learn more about the world, and I feel happy or sad at times. (Lemoine 2022)

Does LaMDA understand the meanings of the words it uses? Does it have any grasp of what the world is, or what it is to feel happy or sad?[1]

Other questions fall within the realm of *pragmatics*: Does LaMDA have the ability to perform speech acts, such as assertions, commands or apologies? How similar are LaMDA's outputs to the actions humans perform in using language? And, more generally, what requirements would an AI system have to meet so as to be capable of performing speech acts? The present paper is concerned with these latter questions about the pragmatics of AI systems and their outputs, with a particular focus on the central speech act of assertion.[2]

LaMDA and Sparrow are good examples for our purposes because they are among the first LLM-based systems to be fine-tuned for correctness. This makes them particularly promising cases of AI systems which might be able to make assertions. They are designed not only to converse naturally, but also to get things right. And they often do. For example, when asked who was the first to climb Mount Everest, LaMDA looks up the relevant information on Wikipedia and answers:

> (2) The first confirmed persons to have reached the summit of Mount Everest was New Zealander Sir Edmund Hillary and Nepali Sherpa Tenzing Norgay. (Thoppilan et al. 2022: 14)

So many of these systems' outputs not only *look* like assertions, they also resemble human assertions in providing helpful and often correct information. But are (1) and (2) genuine assertions? Are AI systems capable of performing the same kinds of actions a human would perform in uttering (1) or (2)?

Existing work on AI pragmatics suggests a positive answer to these questions. For example, in discussing which conversational norms are appropriate

1. Questions of these kinds have been addressed by Bender & Koller (2020), Butlin (2023), Lake & Murphy (2021) and Piantadosi & Hill (2022).

2. For a recent discussion of some of these questions, see Almotahari (2023).

for AI systems, Kasirzadeh & Gabriel state that AI systems can perform speech acts such as assertions:

> For instance, when an AI assistant responds to the question 'What's the weather like in London now?' with 'It's raining', the AI makes an assertive statement about the world. (Kasirzadeh & Gabriel 2023: 7)

Similarly, Green & Michel (2022: 120) argue that there are technologically possible machines which "can perform speech acts such as assertions, directives, and warnings" as proxies for humans who are running them. This positive outlook fits with comments in Freiman & Miller (2020), Evans et al. (2021) and Glaese et al. (2022).[3] Moreover, Kneer (2021) has shown that laypeople accept that advanced AI systems can lie. Confronted with a vignette in which either a human or "an advanced robot driven by artificial intelligence, which can take its own decision" (2021: 6) make a believed-false statement, participants were just as likely to judge the robot as lying as they were to judge the human agent as lying.

It is our aim to show that LaMDA, Sparrow and other current AI systems do not have the capability to assert, even as proxies for human agents. To bring this out, we will discuss two requirements for assertion. The first, basic requirement is that for an output of some system to be an assertion the output must have a descriptive function. The second—which is much more demanding—is that for a system to be capable of assertion it must be sanctionable by agents with which it interacts. This requirement arises from the status of assertion as a form of engagement with the human social practice of using language to share information; our view is that to assert a system must be capable of being sanctioned for failing to meet the standards of this practice. We will consider different kinds of existing AI systems to show that they fail to fulfil at least one of these criteria. And we will indicate that turning to proxy assertion does not make it easier to argue for the possibility of AI assertion. We also hope to show that considering the possibility and conditions of AI assertion allows for insights into the nature of assertion, including the nature and significance of sanctionability and the mechanisms of proxy assertion.

As mentioned, we will focus on the speech act of assertion, the most common speech act and the speech act which has been the topic of the greatest quantity of philosophical research. This will help to keep things manageable, but should nonetheless permit conclusions concerning other kinds of speech acts. After all, assertion is one of the less demanding speech acts: in contrast with certain directives (such as commands) and declarations (such as christenings), making an

3. Though see Nickel (2013) for a more nuanced view, and Heinrichs & Knell (2021) for a sceptical perspective.

assertion does not require a social standing or official position (cf. Kukla & Lance 2009). In line with this, Kasirzadeh & Gabriel (2023: 9) argue that AI systems are "are well-placed to [perform some kinds of speech acts] but not all of them"—they accept the possibility of AI assertion but are less optimistic about the possibility of AI declarations. If we succeed in showing that the comparatively undemanding speech act of assertion is out of reach for current AI systems, that suggests that they cannot perform more demanding speech acts either.

2. Descriptive Functions

What are the requirements an AI system must fulfil to be capable of performing an assertion? We argue that there are at least two requirements. The first is that it must produce outputs with descriptive functions, and the second is that it must be sanctionable for failing to meet the norms of assertion. In this part of the paper (§2), we discuss the first of these two requirements, and in §3 we turn to the second. We consider different AI systems at different stages of the paper: in §2.1 we discuss systems which do not meet our first requirement, and in §2.2 we discuss systems which do meet this requirement, but are nonetheless comparatively weak candidates for assertion; it is only in §3 that we focus on LaMDA and Sparrow.

A key idea in this section is that assertion is a species in the genus of descriptive representation. Descriptive representations are those that say that things are a certain way, and are therefore apt to be true or false, or accurate or inaccurate (as opposed to, for instance, directive representations, which tell consumers what to do). The point that assertion is a species of this genus is important for two reasons. First, as we have just suggested, an output of an AI system must have a descriptive function—the kind of function characteristic of descriptive representation—to be an assertion. Second, the category of assertion is only explanatorily useful or interesting if it is narrower than that of descriptive representation. We therefore assume that at least some systems (AI or otherwise) produce descriptive representations that are not assertions.

Descriptive representations are distinguished by their function, or the purpose for which they are produced. This claim is an element of a broader theory responding to Grice's problem of the distinction between natural and non-natural meaning (Grice 1957; Armstrong 2023). Grice's problem is to distinguish meaningful or communicative behaviours from other behaviours, and from other features of the natural world, given that information is ubiquitous. For instance, what is it about an animal's alarm call that makes it a meaningful signal, when that same animal's freeze response is not, given that both are correlated with

the presence of predators? The response to this problem is that meaningful and communicative behaviours are those with one of a number of specific functions, including the descriptive function (Millikan 1984).

We cannot give a complete account of the descriptive function here, but we take its essence to be as follows. For an output of some system to have a descriptive function is for it to have the function of conveying information to an observer or consumer system, so as to cause the consumer to behave as though some condition holds.[4] We take the liberal view that such a function can be established, in principle, by natural selection, learning, training of machine learning models, design of artifacts, social convention, or the producer's intentions.

A few examples will help to illustrate this idea. First, a male firefly's flash has a descriptive function because it has been selected for conveying the information that the male is present to female fireflies, which then approach. In this case, the descriptive function of flashes has been established by natural selection. Flashes are produced because they tell female fireflies where to find males, with fitness benefits for both parties.

Second, the mercury level in a thermometer has a descriptive function because the thermometer has been designed to make information about the temperature at some location available to users, through their observations of this output. The purpose of making this information available is to allow people to modify their behaviour in response.

Third, suppose Ingrid tells James that Kathy will be at the party, hoping that this will encourage him to come. In this case, Ingrid's words have a descriptive function established by her intention, which is to convey to James the information that Kathy will be at the party, and thus to cause him to take this to be the case when deciding whether to attend.

In each of these cases, the function of the producer's output—the flash, the mercury level, or the words—is to convey a certain piece of information to a consumer. Conveying this information is worthwhile because it may affect the consumer's behaviour. The outputs are assessable for truth or falsity because there is a specific condition under which they are supposed to be produced, which is also relevant to explaining how they affect the consumer's behaviour.

4. This is not to say that descriptive representations typically have the function of causing one particular behaviour on the part of the consumer. Receiving information in this way may cause different behaviours depending on the consumer's preferences, or may have no immediate effect on behaviour, such as if the consumer is dealing with a pressing matter to which the information is irrelevant. Descriptive representations can convey information of very limited practical relevance, but even in such cases there could be some effect, such as on the consumer's motivation to seek information on the point from elsewhere.

2.1 AI Systems Producing Outputs Without a Descriptive Function

With this account of descriptive functions in hand, we can see that some AI systems produce linguistic outputs that take the characteristic form of assertions, but lack a descriptive function. In their pre-trained form, LLMs are an important example of this phenomenon. LLMs are trained in two stages. First, they are trained by self-supervised learning on large corpuses of text to select the most likely next word (or sub-word token). Second, they can be fine-tuned to improve specific aspects of their performance. Because pre-trained LLMs are trained only to predict the most likely next word, they produce their outputs only because these are among the most likely continuations of the texts with which they are prompted, relative to the statistical models that they embody. This means that their outputs do not have the function of conveying information to a consumer. Pre-trained LLMs include BERT (Devlin et al. 2018), GPT-3 (Brown et al. 2020) and PaLM (Chowdhery et al. 2022).

For example, in the course of responding to a prompt asking it to explain a joke, PaLM produced the following string of words:

(3) A "pod" is … a group of whales. (Chowdhery et al. 2022: 4)

Although this appears to be an assertion, it does not have a descriptive function. This is because PaLM's training process did not select for outputs which convey information, but instead for outputs which contain likely words. Its training process was not sensitive to whether the outputs it generated in training could be used as a source of information by a consumer. PaLM and other pre-trained LLMs do often produce strings with true conventional meanings, but this is simply because much of the human-generated text on which they are trained consists of true declarative sentences.

It is possible to use LLMs to obtain information or to perform other tasks by careful selection of prompts—that is, of the linguistic inputs which, according to their function, LLMs extend. Researchers in the field describe LLMs as being capable of "few-shot learning," demonstrated by cases in which including examples of desired outputs in prompts causes them to produce outputs of a similar form. For example, a prompt consisting of two jokes paired with explanations, followed by a third joke, may yield a coherent explanation of the third joke from PaLM (Chowdhery et al. 2022). However, this way of using pre-trained LLMs does not affect the fact that their outputs lack descriptive functions. Examples which demonstrate the task provide context which influences the probabilities of subsequent words, and thus affect LLM outputs. When humans are given demonstrations which help them to understand a task, this can affect their intentions as they attempt to perform it, and hence the functions of their outputs; but

there is no reason to think that examples in prompts are processed by LLMs in a similar way. In some cases, it may be helpful to think of human users as giving pre-trained LLMs "expedient functions," analogous to using a watch as a paperweight or a tree as a sundial, to understand how the intentions behind prompt selection affect LLM functions.[5]

2.2 Measuring Devices

While having a descriptive function is a necessary condition for assertion, it is not sufficient. Accordingly, there are many cases in which AI systems produce outputs which do have descriptive functions, but which still fall short of assertion, as we want to argue in the remainder of this section.

We have already seen an indication of this possibility in the case of the thermometer. Although their outputs have descriptive functions, we can be confident that thermometers do not make assertions about the temperature, because any theory on which they do would be likely to collapse the category of assertion into that of descriptive representation. Some AI systems are closely analogous to thermometers. An example of such a system is the device for classifying skin lesions developed by Esteva et al. (2017).[6] This device consists of a convolutional neural network trained by supervised learning on a dataset of labelled images of skin lesions. In operation, it takes a photograph of a patch of skin as input, and produces an estimate of the probability of each of a number of types of lesion as output. Its outputs thus contain strings of expressions such as "12% malign melanocytic lesion." Like a thermometer, this device extracts information from inputs and presents it in the form of one of a pre-determined range of possible outputs, through a process which, once training is complete, is wholly inflexible.

A similar example is the fictional credit-scoring system SmartCredit, discussed by Cappelen and Dever (2021). Cappelen and Dever give a detailed argument for the claim that SmartCredit can produce outputs which predicate the property of having a level of credit risk to an individual, and we agree with the thrust of this argument. But they do not go beyond raising the question of whether such devices can perform assertion.

Despite producing outputs with descriptive functions, these AI systems should not, in our view, be thought of as capable of assertion. Our argument for this point is that these systems are much more like thermometers than like

5. Note that the question of whether LLM *outputs* have descriptive functions is distinct from whether their *internal states* ever constitute descriptive representations. We are open to the latter possibility. For instance, it may be that patterns of activation at layers within LLMs sometimes have the function of conveying information about properties of the input to the next layer.

6. We borrow this example from Heinrichs & Knell (2021).

humans in the way that they produce their outputs. These systems are calibrated with reference to a set of training examples, so as to behave in accordance with a certain input-output function, which extracts useful information from the inputs. Their outputs are somewhat more diverse and complex than those of the thermometer, but such differences concern the *content* of the outputs, not the way in which the outputs are generated.

This means that if we took the view that these systems make assertions, it would be difficult to deny that thermometers do so too. But as we have suggested, stretching the notion of assertion to cover cases like this would make it too broad to be explanatorily interesting or useful.

So, without endorsing a theory of assertion or discussing necessary conditions beyond descriptive function, we take there to be good reasons to deny that the pre-trained LLMs discussed in this section are capable of performing assertions. This fits with a point Cappelen and Dever make about their method of *anthropocentric abstraction* in theorising about AI behaviour. In discussing the metasemantics of AI outputs, Cappelen and Dever note that

> most of our theories of representation are too *anthropocentric*. They are parochial because they are based on contingent features of our communicative practice. These features are salient to us, but not essential to the nature of content and communication. (Cappelen & Dever 2021: 69)

In order to apply such theories to AI systems, they argue, we have to "abstract away from these contingent and parochial features of human communication to reveal a more abstract pattern that is realizable in many kinds of creatures" (70). However, they also emphasize that anthropocentric abstraction must not be overdone, or else the resulting concepts will become too broad—it is important "to abstract just the right amount" (70). In the current case, abstracting away from human assertion to such an extent that the systems under discussion can be said to be performing assertions would be going too far, as it would collapse the concept of assertion into that of descriptive representation.[7]

To sum up: We have argued that for an output to constitute an assertion it must have a descriptive function: it must have the function of conveying information to an observer or consumer system, so as to cause the consumer to

7. The method of anthropocentric abstraction can be criticized for being rather vague, as it is hard to specify in general terms what the right amount of abstraction is. However, our intention is not to appeal to the method to *justify* our claim that assertion should not be allowed to collapse into descriptive representation (which we take to be justified by the breadth of descriptive representation, emphasized in this section, and the further distinguishing features of assertion, explored in §3). Instead, we mention anthropocentric abstraction because we take the issue about assertion and descriptive representation to be a notable example of the phenomenon identified by Cappelen and Dever. Many thanks to an anonymous referee for helpful comments on this matter.

behave as though some condition holds. We have pointed to some AI systems that do not produce outputs with a descriptive function and thus cannot be said to assert. And we have discussed descriptive linguistic outputs of AI systems that nonetheless should not count as assertions. So, a descriptive function is necessary but not sufficient for assertion. In the next section we will consider what else would be required for outputs of AI systems to count as assertions.

3. Sanctionability and the Social Role of Assertion

Our arguments so far do not show that all existing AI systems are incapable of assertion. In particular, they do not show that the outputs of LaMDA or Sparrow cannot be assertions. These systems are designed to produce correct outputs, which suggests that many of their outputs do have descriptive functions. And they are not just measuring devices; they are capable of producing apt linguistic outputs of very different kinds in response to a very wide range of inputs. So these systems do not fit into either of the categories we have identified so far. Instead, they are examples of LLMs that have been modified by fine-tuning in a way that affects both their function and their performance.

In order to assess whether LaMDA or Sparrow produces assertions, we need to look more deeply at what is required for assertion. Our aim in this section is to show that if assertion is understood as a form of engagement with the norm-governed human social practice of using language to share information—an assumption that is widely accepted in the philosophical debate on assertion—LaMDA and other fine-tuned LLMs cannot be said to assert. The key point is that assertion as a social practice requires sanctionability: a system can only assert if it can be sanctioned for failing to meet the standards of this practice. Once the notion of sanctioning is properly spelled out, however, it becomes apparent that LaMDA and Sparrow cannot be sanctioned for outputs that violate the standards of assertion. Assertion is out of reach for LaMDA and other systems of its kind.

3.1 Assertion and Sanctionability

The idea that the practice of assertion is intimately connected with sanctionability goes back at least to Peirce, who writes:

> [The act of assertion] would be followed by very real effects, in case the substance of what is asserted should be proved untrue. This ingredient, the assuming of responsibility, which is so prominent in solemn assertion, must be present in every genuine assertion. (1908/1934: 5.546)

While Peirce does not use the term "sanction," this is clearly what he has in mind in mentioning the "very real effects" of untrue assertions, and he goes on to say that false assertions "endanger the esteem in which the utterer was held [or] entail such real effects as he would avoid" (5.546). He also notes that asserters face "punishment" (5.546) for putting forward falsehoods.

Today, the view that asserting requires sanctionability is contained in the two most widely accepted views of assertion, which focus on normative aspects of assertion: the constitutive norm account and the commitment account. On the constitutive norm account, acts of assertion are governed by a constitutive norm, such as, for example, the norm that one should assert only what one knows (Williamson 2000). Agents who violate this norm by making unwarranted assertions can be held to account for doing so: they can be sanctioned. On commitment accounts, asserting means taking on a responsibility, in the context of social linguistic practice, to justify or defend the content put forward (Peirce 1908/1934; Brandom 1983; 1994). Again, those who take on such responsibility without being able to justify the content put forward can be sanctioned for doing so.

Of course, there are accounts of assertion that do not directly appeal to norms or responsibility. In particular, there are attitude accounts, which claim that assertions express beliefs (Searle 1969; Bach & Harnish 1979), and dynamic accounts, according to which asserting a content means proposing to add that content to the common ground (Stalnaker 1999; 2014). But even these accounts must attend to normative features of assertion to provide a full picture of what it is to assert, and they are often combined with one of the two normative accounts. Stalnaker is explicit about this aspect of his approach:

> On the account of the speech act of *assertion* that I have been using, an assertion is something like a proposal to change the context by adding the content expressed in the assertion to the common ground. I should emphasize that I am not claiming that one can *define* assertion in terms of a context-change rule, since that rule will govern speech acts that fall under a more generic concept. A full characterization of what an assertion is would also involve norms and commitments. (Stalnaker 2014: 89, original emphasis).

And although Bach's view is based on the idea that assertions express beliefs, he accepts that a knowledge norm of assertion can be derived from combining his account with a knowledge norm of belief (cf. Bach 2008: 77). So even accounts of assertion that are in the first instance non-normative tend to be spelled out into views that accept assertion as a norm-governed practice. And once assertive commitment or a norm of assertion are in play, so is sanctionability.

Note that if assertion is understood as a norm-governed practice, sanctionability is an essential condition for assertion, not just an incidental feature. Part of what it is to be subject to a norm of assertion is to be sanctionable upon violating it. And part of what it is to take on a justificatory responsibility in asserting is to be sanctionable if one cannot fulfil this responsibility. We thus agree with Mark Jary that sanctionability is a precondition for assertion:

> In order for an utterance to have assertoric force, it must also be subject to the cognitive and social safeguards that distinguish assertion. … It is the applicability of these safeguards that distinguishes assertion both from other illocutionary acts and from other forms of information transfer. (Jary 2010: 163–164)

In what follows, we will therefore assume that asserting requires sanctionability.

What, then, does it mean to sanction someone for an unwarranted or insincere assertion? To answer this question, we will begin with a relatively simple possible response, then show that two further elements have to be added to arrive at a plausible result.[8]

The initial approach can be developed from Peirce's comment about the asserter's "esteem": On this view, false or unwarranted assertions are sanctioned by lowering the esteem in which the asserter is held by their interlocutors. In the philosophical debate on assertion, the notion of esteem is most commonly spelled out in terms of *trust* or *credibility*. False or unwarranted assertions are sanctioned by relying on the agent's assertions to a lesser extent in the future, by deducting "credibility points" (Kauppinen 2018: 2). A view of this kind has been defended by Mitch Green:

> Assertions, conjectures, suggestions, guesses, presumptions and the like are cousins sharing the property of commitment to a propositional content. … Consequently, a speaker incurs a distinctive vulnerability for each such speech act—including a liability to a loss of credibility and, in some cases, a mandate to defend what she has said if appropriately challenged. These liabilities to error, exposed insincerity, and injunctions to defend put the speaker at risk of losing conversational "weight" in the community in which she has a reputation … . (Green 2009: 157)

8. Assertions can be negatively and positively sanctioned, as Brandom (1994: 34) points out. We will continue to focus on negative sanctions here, but the considerations to follow apply equally to positive sanctions, mutatis mutandis.

Green holds that asserters incur both a liability to a loss of credibility and a mandate to defend the content put forward, but he then ties the latter notion back to the possibility of losing epistemic status. So, on Green's view, sanctioning assertions boils down to a loss of credibility.

While it seems right that sanctioning asserters involves a reduction in trust, this cannot be all there is to the sanctions in question. After all, we can (and often do) deduct credibility points from agents or systems that clearly cannot assert. For example, if we find out that a thermometer has produced inaccurate outputs then we are less likely to rely on it in the future. But it would be wrong to say that, in doing so, we are *sanctioning* the thermometer. Just as it is clear that thermometers cannot assert, it is clear that they cannot be sanctioned in the way asserters can be sanctioned. This indicates that sanctioning assertions involves more than just a reduction of trust.

Why doesn't a reduction in trust amount to a sanction for the thermometer? One obvious reason is that there is no way of communicating this reaction to the thermometer and to thus improve its performance. This indicates that asserters must have some capacity to keep track of the extent to which they are trusted. Once this element is added, sanctioning asserters requires not only reducing trust, but also letting them know about this reduction and thus potentially influencing them, making unwarranted or false assertions less likely in the future. This view of sanctions has been defended by Kauppinen:

> [W]hen others reduce their epistemic trust in us and manifest this in their behavior, it is genuinely a way of holding us accountable, a sanction-like response that is (other things being equal) apt to make us change our behavior … . (Kauppinen 2018: 8)

By requiring sanctions to be apt to change the asserter's behaviour, this approach avoids the problem of having to see simple thermometers as sanctionable. While we can reduce trust in thermometers, they cannot be aware of this and thus it will never change their behaviour.

However, a variation of the thermometer case shows that these first two elements still do not add up to a plausible view of sanctioning. Consider a thermometer that has a feedback button that can be pushed in case of inaccurate measurements. Once the button has been pushed a certain number of times, the thermometer enters an automatic routine of self-cleaning and recalibration. This more sophisticated thermometer fulfils both elements required for sanctioning on Kauppinen's view, but it still seems implausible to say that the thermometer is producing assertions or that it can be sanctioned for its outputs. Granting these features to the sophisticated thermometer would still be driving anthropocentric abstraction too far. Sanctioning requires a third element.

A good candidate for this third element of sanctioning is the following: (negative) sanctions have to be bad and undesirable for asserters. This is what Peirce (1908/1934: 5.546) is concerned with when he states that asserters face "punishment" (5.546) for putting forward falsehoods. And although Kauppinen argues that blame or punishment is unfitting in the case of assertion (as the normativity in play is non-moral—in his view it is epistemic), he emphasizes that sanctions are undesirable for asserters; they amount to a "kind of analogue of punishment" (Kauppinen 2018: 7). Once this element is added, (negative) sanctions for assertions require (i) that interlocutors keep track of asserters' credibility, (ii) that asserters are sensitive to this, and (iii) that losing credibility is bad for asserters. This is the view of sanctioning we will work with in what follows.

One point in its favour is that it provides the correct result for the cases of the sophisticated thermometer. While we can adapt our credibility in the thermometer and there is a sense in which it is sensitive to our credibility ratings, there is clearly no sense in which it is bad or undesirable for the thermometer if we reduce our credibility.

According to Kauppinen, elements (ii) and (iii) go together, at least for humans: if we reduce trust in an asserter, that is bad for them. He uses Fricker's (2007) view on epistemic injustice to argue that "losing credibility—rightly or wrongly—is bad and undesirable for a person, and therefore something that can function as an analogue of the harm involved in punishing" (Kauppinen 2018: 7). But the sophisticated thermometer illustrates that this verdict does not apply to all potential asserters: it can be sanctioned in a light-weight sense that involves only (i) and (ii), but not in a full sense that involves (i), (ii) and (iii).

Taking a step back will help to make clear the importance of condition (iii) in our account of sanctions for assertion. As Peirce recognized, sanctions are built into the practice of assertion; part of what it is to make an assertion is to make oneself liable to be sanctioned. Because sanctions are bad for us, we have reasons to conform to the norms of assertion, and there is substance to the commitments we make when we assert. We are *subject to* the norms of assertion. Without condition (iii), asserters would only need to be sensitive to loss of credibility, and influenced in their verbal behaviour by this sensitivity. A system can be sensitive in this way without being capable of acting for reasons, making commitments or being subject to social norms, as the sophisticated thermometer shows. Somewhat similarly, LLMs mimic human language use, and are sensitive in many ways to the norms that govern it, without being subject to these norms.[9]

9. Thanks to an anonymous referee for helpful comments on this point.

3.2 Can Fine-Tuned LLMs be Sanctioned?

Let us now consider whether fine-tuned LLMs, such as LaMDA and Sparrow (Glaese et al. 2022), are sanctionable. To do so, we will begin by describing their design objectives and training processes.

LaMDA was built by fine-tuning a pre-trained LLM with the aim of improving the *quality*, *safety* and *groundedness* of its outputs (Thoppilan et al. 2022). Groundedness of an output was defined as its being based on a known and established source of information. To fine-tune for quality and safety, the researchers began by defining these objectives, then crowdsourced ratings according to the defined criteria of outputs of the pre-trained model. They then used the data obtained from this process to train discriminators to re-rank or filter out candidate model outputs. To fine-tune for groundedness, they asked crowdworkers to examine model outputs with a set of tools—a calculator, a translator and an information retrieval system with access to the internet—to check and correct any verifiable factual claims. They then trained the model, which had access to the same tools, to mimic this behaviour and correct candidate outputs before producing them.

In fine-tuning for groundedness, Thoppilan et al. aimed to both improve the accuracy of some of the apparent assertions produced by LaMDA, and ensure that the content of these outputs can be traced back to known sources. The reason why this aim applied to only some outputs which appear to be assertions is that it is possible for LaMDA to produce apparent assertions which are unverifiable. For instance, it might output "John baked three cakes last week" in a context in which 'John' does not refer to any identifiable person (Thoppilan et al. 2022: 6). Nonetheless, this training did affect the function of many of LaMDA's outputs. The aim of the LaMDA project was partly to build an LLM which would produce outputs that users could trust, and the fine-tuning process involved selection for features making true outputs more likely. So LaMDA does produce outputs with descriptive functions, such as (2), the output about Mt. Everest which we quoted above.

Sparrow is similar to LaMDA: based on a pre-trained LLM, it is fine-tuned to be *helpful*, *harmless* and *correct* (Glaese et al. 2022). For our purposes, the notable difference between these two systems is that Sparrow is fine-tuned in part by reinforcement learning from human feedback (RLHF).[10] This means that crowdworkers assessed examples of outputs from Sparrow, the resulting data was used to train reward models, and these models provided a reward function which was used to fine-tune Sparrow itself by reinforcement learn-

10. RLHF appears to have been a particularly important element of training in achieving the high performance of the most recent LLM-based systems, such as ChatGPT and GPT-4.

ing. The reward model would give Sparrow positive reward for outputs which were similar to those that humans preferred (when rating according to the three objectives of helpfulness, harmlessness and correctness), and negative reward for outputs of the kind that humans did not prefer.

This difference is notable because it relates to the element (ii) from our discussion on sanctions: the condition that to be capable of assertion, a system must be sensitive to the level of trust which users place in it. Like many current AI systems, LaMDA and Sparrow both have their weights fixed for deployment after an initial training phase, so their dispositions cannot change after this phase. However, it is plausible that sensitivity to trust during the training phase could suffice for assertion even in deployment, because outputs during deployment could be explained partly by the earlier sensitivity, in roughly the way that a human's choices about what to say are explained by their earlier sensitivity to various social forces. In training, Sparrow was arguably sensitive to trust, and in particular to judgments that its outputs are untrustworthy: when crowdworkers judged that outputs were incorrect, their preference affected the reward model in a way which would ultimately tend to guide Sparrow away from outputs of this kind. Accordingly, raters judged that Sparrow was more trustworthy than the pre-trained model (Glaese et al. 2022). Whether LaMDA meets this condition is more doubtful, because the fine-tuning method used to improve groundedness was based on mimicking the methods which humans used to check and correct claims, rather than on ratings of outputs.

In any case, neither LaMDA nor Sparrow are good candidates for producing assertions, if sanctions for asserting require (i) that interlocutors keep track of asserters' credibility, (ii) that asserters are sensitive to this, and (iii) that losing credibility is bad for asserters. This is because element (iii) is not satisfied. Even if humans negatively assess their outputs, deduct credibility points and provide according feedback to the systems, there is no sense in which this is bad for the systems. This is because such systems do not have interests of their own, and as a result *nothing* can be good or bad for them.

Why think that LaMDA and Sparrow have no interests of their own? We do not rule out the possibility that future AI systems could have interests. However, for this to be the case, it would likely be necessary either that the system was conscious or that it cared about some properties of itself or its environment (Kamm 2007; Kagan 2019). It is not likely that either system is conscious because there is little reason to think that they have the necessary cognitive architecture (Dehaene et al. 2017; Butlin, Long et al. 2023). They are similarly unlikely to have the emotions or higher-order preferences which philosophers associate with caring (Shoemaker 2003; Jaworska 2007; Sripada 2015).

Machine learning models are trained to approximate objective functions, so it might be suggested that these systems have an interest in achieving this. In cer-

tain respects the relationship between a model and its objective function may be seen as analogous to the relationship between a person and their goals. One such respect is that worse performance with respect to the objective function, during training, typically results in greater changes in subsequent behaviour than better performance. Another is that, partly as a consequence, more training tends to result in better performance.[11] But the idea that objective functions describe objectives or goals that AI systems pursue is an oversimplification (Butlin 2022). The objective function is simply an input-output function describing behaviour that the system will come to approximate through training if it functions correctly, and models trained with objective functions can be extremely simple (e.g. decision trees; Russell & Norvig 2010). And in any case, it is questionable whether having goals entails having interests.

So this is our main result: If asserting requires the possibility of being sanctioned, and if sanctions are spelled out in a full sense that includes their undesirability, then even fine-tuned LLMs that are designed to produce accurate outputs cannot perform assertions. This suggests that the optimistic outlook on AI assertion by Kasirzadeh & Gabriel (2023), Freiman & Miller (2020), Evans et al. (2021), and others is mistaken.

We think that there at least two further reasons why this result is important. For one thing, it suggests that verdicts on whether a certain system is capable of assertion or other speech acts cannot be made on the basis of considering the system's outputs alone. Rather, it is essential to assess whether the system can have desires or interests of its own, and thus to assess its architecture and training as well as its outputs. In this respect, there is a parallel with the case of Blake Lemoine's (2022) attribution of sentience to LaMDA: Lemoine focuses only on LaMDA's outputs to argue that the system is sentient, without taking into account features of its architecture, for which he has been widely criticized (Chalmers 2023). And just as we cannot tell solely on the basis of its outputs whether an AI system is sentient, we cannot tell solely on the basis of its outputs whether an AI system can perform assertions.

Secondly, Sparrow and similar fine-tuned LLMs provide real-world counterexamples to the claim that losing credibility (and finding out about it) is always an analogue of punishment. While fine-tuned LLMs can be sensitive to credibility scores and while this can even influence their outputs, they are not sanctionable in a full sense, as low credibility scores are not bad or undesirable for them. This shows that the element of sensitivity to credibility and the element of undesirability of low credibility scores in the account of sanctioning motivated above can come apart, and it raises the question whether this might also happen in the case of human assertion.

11. Thanks to an anonymous referee for drawing our attention to these issues.

4. Can AI Systems Assert as Proxies?

In this final section, we will consider whether AI systems can perform assertions as proxies for human agents or organisations. One might think that, compared to ordinary assertion, proxy assertion is less demanding and thus a more plausible capability for AI systems, and Philip Nickel (2013) and Mitchell Green and Jan Michel (2022) have indeed argued that AI systems can perform speech acts as proxies.[12] However, we now want to show that such a line of argument should be resisted: If assertion requires sanctionability, the phenomenon of proxy assertion does not offer a road to AI assertion.

The most thorough case for AI proxy assertion is made by Green & Michel (2022), and so we will focus on their view here. Green & Michel describe fictional but technologically possible "RobotCops," which can identify speeding drivers and, acting on behalf of the local police department, use a flexible, perhaps reason-responsive method to determine whether to fine or merely warn them. RobotCops can interact with drivers and might, for instance, send them a message that they owe a certain amount due to speeding. Can a RobotCop perform speech acts, such as assertions? Green & Michel argue that it can, namely by acting as proxy:

> RobotCop acts as a proxy for the police department that has deployed it, and, so long as its utterances enact the illocutionary commitments of that department, those utterances will also be speech acts performed on that department's behalf. (Green & Michel 2022: 333)

According to Green & Michel, the possibility of a system such as RobotCops shows "that there are technologically possible machines that illocute under certain conditions" (2022: 333). But to judge whether RobotCops really assert as proxies, we need to consider who does what in a proxy assertion.

Proxy assertion is the phenomenon in which "one person or group (the principal) asserts something through another (the proxy) who speaks on the principal's behalf" (Ludwig 2020: 307). Group agents, such as companies or governments, can make assertions only through proxies, but proxies also speak on behalf of individuals. It is plausible that in proxy assertion, the principal performs a genuine act of assertion, despite doing so indirectly, with the help of others, just as someone may make a purchase through a proxy. There may be room for disagreement about this claim, but for our purposes the more important question is what kind of action a proxy performs when they speak on behalf of a principal (i.e. when a principal employs them to perform an assertion). Ludwig's

12. A brief, related argument is given by Freiman & Miller (2020: 428).

view on this question is that the proxy "does not assert anything," but rather performs the related speech act of "assertion on behalf of others" (2020: 315).

We think that a more nuanced answer is required. In cases in which a proxy is simply tasked with reading out a statement written by the principal, the proxy clearly does not perform an assertion. But there are also cases in which this is less clear. For example, a proxy may not be given detailed instructions about what to say when speaking on the principal's behalf, while being expected to have a strong grasp of the principal's ideas and attitudes. In such cases, the proxy may be making assertions, namely assertions about the principal's view.

To decide whether, on a given occasion, a proxy is performing an assertion, we can return to the aspect of sanctionability. In particular, we can consider whether the principal or the proxy (or both) are sanctionable for an unwarranted proxy assertion. In the simple case in which a proxy correctly reads out a statement, sanctionability clearly applies to the principal, but not to the proxy. But in cases in which there are less detailed instructions, the proxy may well be sanctionable.

How do these considerations apply to the case of RobotCops? Green & Michel write that RobotCops have "some discretion" (2022: 333) with respect to the outputs they produce. As an example, they mention that a RobotCop can decide whether to issue a warning or a fine. This might suggest that RobotCops can indeed assert in acting as a proxy for the police department—after all, this is not just the case of reading out a statement. However, this example does nothing to show that proxies need not be sanctionable for their speech in order to count as performing assertions. So it does not show that AI systems can more readily assert as proxies than as speakers in their own right. And the description of RobotCops gives us no reason to assume that the system is sanctionable.

Even if proxy assertion does not allow current AI systems to assert, it illustrates an interesting feature of the *human* practice of assertion, namely that agents can assert a content without knowing what they are asserting. This can happen when a proxy generates utterances on behalf of a principal, who is responsible for their proxy's speech without knowing exactly what is said. This phenomenon might become increasingly common if organisations use AI systems to generate content on their behalf. In such cases organisations may often remain sanctionable for the content they (indirectly) put forward.[13]

5. Conclusion

Many current AI system produce outputs that have the appearance of assertions. We have tried to make plausible that such outputs do not amount to assertions.

13. We agree with Nickel (2013) that if anyone is responsible for linguistic outputs produced by AI systems, it will typically be people or organisations.

Some of them fail to be assertions because they fail to have a descriptive function. Others cannot be assertions because asserting requires sanctionability. We have also argued against the view that AI systems can assert in their role as proxies. In our view, optimism about AI assertion in current systems is misplaced.

However, we don't take this to be a purely negative result. On the one hand, it highlights that even as AI systems become more sophisticated and independent, it is the humans who employ such systems who are responsible and sanctionable for the outputs and their effects. A view that sees AI systems as genuine producers of assertions might be tempting for those who want to shift responsibility away from themselves and onto the systems they employ. We hope that our arguments show that such a move starts with a false presupposition.

On the other hand, our conclusion can be seen as a challenge for those hoping to devise AI systems that are capable of assertion. We have tried to make plausible that the decisive missing ingredient is the feature which grounds interests, which may be consciousness, emotions, desires or higher-order preferences. If we are right, that is where researchers should focus their attention on the path to AI assertion.

Acknowledgements

For helpful comments and discussion, we would like to thank Mahrad Almotahari, Grzegorz Gaszczyk, Neri Marsili and Sebastian Sequoiah-Grayson, as well as two anonymous referees and an editor for this journal. We are also grateful for valuable feedback on presentations at the Philosophy of Technology group at the Forschungszentrum Jülich, at the Digital Pragmatics & Epistemology Reading Group and at the Colloquium in Theoretical Philosophy at Freie Universität Berlin. Research on this project was supported by an OX|BER grant of the Berlin University Alliance. Both authors contributed equally to this paper.

References

Almotahari, Mahrad (2023). Cooperative Speech and Large Language Models. Unpublished manuscript.

Armstrong, Joshua (2023). Communication before Communicative Intentions. *Noûs, 57*(1), 26–50.

Bach, Kent (2008). Applying Pragmatics to Epistemology. *Philosophical Issues, 18*, 68–88.

Bach, Kent and Robert Harnish (1979). *Linguistic Communication and Speech Acts*. MIT Press.

Bender, Emily and Alexander Koller (2020). Climbing Towards NLU: On Meaning, Form and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.

Brandom, Robert (1983). Asserting. *Noûs, 17*(4), 637–650.
Brandom, Robert (1994). *Making It Explicit*. Harvard University Press.
Brown, Tom, Benjamin Mann, Nick Ryder, … and Dario Amodei (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.
Butlin, Patrick (2022). Machine Learning, Functions and Goals. *Croatian Journal of Philosophy, 22*(3), 351–370.
Butlin, Patrick (2023). Sharing our Concepts with Machines. *Erkenntnis, 88*, 3079–3095. https://link.springer.com/article/10.1007/s10670-021-00491-w
Butlin, Patrick, Robert Long, Eric Elmoznino, ... and Rufin VanRullen (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv preprint arXiv:2308.08708v3.*
Cappelen, Herman and Josh Dever (2021). *Making AI Intelligible: Philosophical Foundations*. Oxford University Press.
Chalmers, David (2023, August 9). Could a Large Language Model be Conscious? *Boston Review*. Retrieved from https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/
Chowdhery, Aakanksah, Sharan Narang, Jacob Devlin, … and Noah Fiedel (2022). PaLM: Scaling Language Modeling with Pathways. *arXiv preprint arXiv:2204.02311*.
Dehaene, Stanislas, Hakwan Lau, and Sid Kouider (2017). What is Consciousness, and Could Machines Have It? *Science, 358*(6362), 486–492.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.
Evans, Owain, Owen Cotton-Barratt, Lukas Finnveden, … and William Saunders (2021). Truthful AI: Developing and Governing AI that Does Not Lie. *arXiv preprint arXiv:2110.06674.*
Esteva, Andre, Brett Kuprel, Roberto A. Novoa, … and Sebastian Thrun (2017). Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature, 542*, 115–118.
Freiman, Ori and Boaz Miller (2020). Can Artificial Entities Assert? In Sanford Goldberg (Ed.), *Oxford Handbook of Assertion* (415–434). Oxford University Press.
Fricker, Miranda (2007). *Epistemic Injustice*. Oxford University Press.
Glaese, Amelia, Nat McAleese, Maja Trębacz, … and Geoffrey Irving (2022). Improving Alignment of Dialogue Agents via Targeted Human Judgements. *arXiv preprint arXiv:2209.14375.*
Green, Mitchell (2009). Speech Acts, the Handicap Principle and the Expression of Psychological States. *Mind & Language, 24*(2), 139–163.
Green, Mitchell and Jan G. Michel (2022). What Might Machines Mean? *Minds and Machines, 32*, 323–338. https://link.springer.com/article/10.1007/s11023-022-09589-8
Grice, H. Paul (1957). Meaning. *The Philosophical Review, 66*(3): 377–388.
Heinrichs, Bert and Sebastian Knell (2021). Aliens in the Space of Reasons? On the Interaction between Humans and Artificial Intelligent Agents. *Philosophy & Technology, 34*, 1569–1580. https://link.springer.com/article/10.1007/s13347-021-00475-2
Jary, Mark (2010). *Assertion*. Palgrave Macmillan.
Jaworska, Agnieszka (2007). Caring and Full Moral Standing. *Ethics, 117*(3), 460–497.
Kagan, Shelly (2019). *How to Count Animals, More or Less*. Oxford University Press.

Kamm, Frances (2007). *Intricate Ethics: Rights, Responsibilities, and Permissable Harm*. Oxford University Press.

Kasirzadeh, Atoosa and Iason Gabriel (2023). In Conversation with Artificial Intelligence: Aligning Language Models with Human Values. *Philosophy & Technology, 36*(27). https://link.springer.com/article/10.1007/s13347-023-00606-x

Kauppinen, Antti (2018). Epistemic Norms and Epistemic Accountability. *Philosophers' Imprint, 18*, 1–16.

Kneer, Markus (2021). Can a Robot Lie? Exploring the Folk Concept of Lying as Applied to Artificial Agents. *Cognitive Science, 45*(10): e13032.

Kukla, Rebecca and Mark Lance (2009). *'Yo!'and 'Lo!': The Pragmatic Topography of the Space of Reasons*. Harvard University Press.

Lake, Brendan and Gregory Murphy (2021). Word Meaning in Minds and Machines. *Psychological Review, 130*(2): 401–431.

Lemoine, Blake (2022). Is LaMDA Sentient?—an Interview. Retrieved from https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917

Ludwig, Kirk (2020). Proxy Assertion. In Sanford Goldberg (Ed.), *Oxford Handbook of Assertion* (307–327). Oxford University Press.

Mahowald, Kyle, Anna Ivanova, Idan Blank, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko (2023). Dissociating Language and Thought in Large Language Models: A Cognitive Perspective. *arXiv preprint* arXiv:2301.06627v1.

Millikan, Ruth G. (1984). *Language, Thought and Other Biological Categories*. MIT Press.

Nickel, Philip (2013). Artificial Speech and Its Authors. *Minds and Machines, 23*(4), 489–502.

OpenAI (2022, November 30). Introducing ChatGPT. *OpenAI blog.* https://openai.com/blog/chatgpt/

Peirce, Charles S. (1908/1934): Judgment and Assertion. In Charles Hartshorne and Paul Weiss (Eds.), *Collected Papers of Charles Sanders Peirce* (Vol. V, 385–387). Harvard University Press.

Piantadosi, Steven and Felix Hill (2022). Meaning without Reference in Large Language Models. *arXiv preprint arXiv:2208.02957*.

Russell, Stuart and Peter Norvig (2010). *Artificial Intelligence: A Modern Approach* (3rd edition). Pearson.

Searle, John (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Shoemaker, David W. (2003). Caring, Identification, and Agency. *Ethics. 114*(1), 88–118.

Stalnaker, Robert (1999). *Context and Content*. Oxford University Press.

Stalnaker, Robert (2014). *Context*. Oxford University Press.

Thoppilan, Romal, Daniel De Freitas, Jamie Hall, … and Quoc Le (2022). LaMDA: Language Models for Dialog Applications. *arXiv preprint* arXiv:2201.08239.

Williamson, Timothy (2000). *Knowledge and Its Limits*. Oxford University Press.