Cognitive Systems Research 83 (2024) 101174

Contents lists available at ScienceDirect

Cognitive Systems Research

journal homepage: www.elsevier.com/locate/cogsys

Does ChatGPT have semantic understanding? A problem with the statistics-of-occurrence strategy

Lisa Miracchi Titus

University of Denver, United States of America

ARTICLE INFO

Action editor: Matteo Colombo

Keywords:
Large language models
Deep learning
Grounded language learning
Semantic understanding
Artificial intelligence
ChatGPT

ABSTRACT

Over the last decade, AI models of language and word meaning have been dominated by what we might call a *statistics-of-occurrence*, strategy: these models are deep neural net structures that have been trained on a large amount of unlabeled text with the aim of producing a model that exploits statistical information about word and phrase co-occurrence in order to generate behavior that is similar to what a human might produce, or representations that can be probed to exhibit behavior similar to what a human might produce (*meaning-semblant behavior*). Examples of what we can call Statistics-of-Occurrence Models (SOMs) include: Word2Vec (CBOW and Skip-Gram), BERT, GPT-3, and, most recently, ChatGPT. Increasingly, there have been suggestions that such systems have semantic understanding, or at least a proto-version of it. This paper argues against such claims. I argue that a necessary condition for a system to possess semantic understanding is that it function in ways that are causally explainable by appeal to its semantic properties. I then argue that SOMs do not plausibly satisfy this *Functioning Criterion*. Rather, the best explanation of their meaning-semblant behavior is what I call the *Statistical Hypothesis*: SOMs do not themselves function to represent or produce meaningful text; they just reflect the semantic information that exists in the aggregate given strong correlations between word placement and meaningful use. I consider and rebut three main responses to the claim that SOMs fail to meet the Functioning Criterion. The result, I hope, is increased clarity about *why* and *how* one should make claims about AI systems having semantic understanding.

0. Introduction

Over the last decade, AI models of language and word meaning have been dominated by what we might call a *statistics-of-occurrence* strategy: these models are deep neural net structures that have been trained on a large amount of unlabeled text with the aim of producing a model that exploits statistical information about word and phrase co-occurrence in order to generate behavior that is similar to what a human might produce, or representations that can be probed to exhibit behavior similar to what a human might produce.[1]

Examples of what we can call Statistics-of-Occurrence Models (SOMs) include: Word2Vec (CBOW and Skip-Gram), BERT, GPT-3, and, most recently, ChatGPT. They also include multi-modal systems such as DALL-E and DALL-E 2. This paper investigates whether SOMs such as these can plausibly be understood to possess *semantic understanding*, or what is often called *grounded language understanding* in the natural language processing (NLP) and inference (NLI) literatures. Do these systems possess understanding of the meanings of their natural language inputs or outputs? Do their architectures serve as plausible grounds for semantic understanding in humans and perhaps other intelligent non-human animals?[2]

Inspired by the impressive results of this strategy, there has been a recent proliferation of discussion on such questions. Such models have been argued to encode commonsense knowledge (Da et al., 2021), linguistic knowledge (Tenney et al., 2019), and practical knowledge (Huang et al., 2022), and to illuminate the mechanisms underlying our corresponding human capacities (Schrimpf et al., 2021). Researchers have argued that SOMs may be able to replace traditional symbolic "knowledge" structures (or "knowledge bases") used for NLP and NLI applications (Petroni et al., 2019). Increasingly, researchers are questioning whether SOMs shed light on human language understanding and capacities for inference (Bhatia, 2017).

E-mail address: lisa.titus@du.edu.

[1] This work was generously supported by a Fellowship from the National Endowment for the Humanities (FEL-282501-22, Lisa Miracchi).

[2] Many of the models under consideration in this paper are often called Large Language Models (LLMs). Since my objection to these models having semantic understanding will not be due to their size (or the size of their training data) but rather to the focus on statistics-of-occurrence as proxy for semantic information, I introduce the term "Statistics-of-Occurrence Model" here. This clarifies the scope of the objection, and also allows me to also include other semantic representations that are not strictly speaking language models — specifically Vector Semantic Representations.

https://doi.org/10.1016/j.cogsys.2023.101174
Received 2 February 2023; Received in revised form 30 August 2023; Accepted 11 September 2023
Available online 20 September 2023
1389-0417/© 2023 Elsevier B.V. All rights reserved.

Moreover, more and more people, both inside and outside the engineering world, are raising concerns that these systems may even have become capable of understanding and communicating meaningfully themselves. The recent public discussion over whether Google's chatbot LAMDA is sentient, following a (now) former Google engineer's claims to that effect, are but one salient example (Tiku, 2022). However, once taken to this extreme, many researchers balk at the claim that language models understand and communicate similarly to humans, pointing out important differences between SOMs and human understanding (Delcid, 2022). Which position is right?

The increasing prevalence and salience of such discussions, both within academia and the public sphere, make it important to provide a more principled, theoretical understanding of the relationship between SOM and human (or other intelligent) capacities. This paper attempts to make progress toward such understanding.

I will argue that SOMs do not plausibly possess semantic understanding, nor are the mechanisms they use likely to be highly explanatory of semantic understanding in humans and intelligent non-human animals. We should reject the *Semantic Hypothesis*, namely that some form of semantic understanding or grounds thereof can be rightly attributed to SOMs.[3] This is because SOMs plausibly owe their behavioral success merely to the existence of systematic correlations between word co-occurrence statistics and semantic relationships, not to functioning sensitively to semantic relationships in ways that could support claims to genuine semantic understanding. As such, we should prefer what I call the *Statistical Hypothesis*: The impressive, meaning-like behavior of SOMs piggybacks on human capacities for meaningful language use but does not replicate it. SOMs do not themselves function to represent or produce meaningful text; they just reflect the semantic information that exists in the aggregate given the correlations between word placement and meaningful use.

In Section 1, I clarify and focus the discussion by narrowing in on a plausible necessary condition for semantic understanding which I call the *Functioning Criterion*: systems with semantic understanding function in ways that respect the contents of their states and processes, such that their behavior is causally explainable by appeal to those contents. I provide some motivation for this criterion and discuss how it imposes a substantial but plausible and empirically tractable requirement on systems with semantic understanding.

In Section 2, I introduce one prominent kind of SOM that will serve as an example for our discussion — Vector Space Models (VSMs) such as Word2Vec. I explain why we should not directly infer semantic understanding from their *meaning-semblant* behavior (i.e. behavior that *looks* meaningful), but rather acknowledge the potential gap between meaning-semblant behavior and internal functioning that would meet the Functioning Criterion.

The remainder of the paper considers three salient strategies for establishing the Semantic Hypothesis. The first argues that the best explanation of behavioral evidence is in fact that the SOM in question satisfies the Functioning Criterion. Section 3 discusses and rebuts this response by discussing case studies both of VSMs and another key category of SOMs, transformer-based language models such as GPT-3. Although these architectures enable the building of much more sophisticated and powerful models, because they are still optimized for predictive accuracy based on statistics-of-occurrence information, we should take the Statistical Hypothesis to be the more plausible hypothesis.

The second strategy is to claim that *all it is* for a system to function sensitively to semantic properties – and so to satisfy the Functioning Criterion – is for a system to function sensitively to statistics-of-occurrence properties. Section 4 rebuts this response by exploring the Distributional Hypothesis that motivates research into VSMs (and plausibly the statistics-of-occurrence strategy more broadly) and discusses the way in which the legitimate use of VSMs as meaning representations can lead to problematic conflations in this domain. I explain why this approach provides a problematic error theory for *human semantic understanding* and should therefore be rejected.

The third, and most interesting, strategy is to claim that *the way* some SOMs achieve their predictive power is to develop functional sensitivity to semantic properties. In Section 5, I consider suggestions by Potts (2022) to this effect. He advocates using Causal Abstraction Analysis (CAA) to assess whether large neural networks function sensitively to semantics. While I agree that this route is promising, I explore a recent CAA of BERT to argue that it does not currently provide strong evidence for the Semantic Hypothesis, and also to motivate further refinement of the approach. Then, I discuss the recently released and highly impressive ChatGPT. While ChatGPT has some features that distinguish it from these other transformer-based language models, the kind of fine-tuning ChatGPT received does not plausibly support claims that its internal functioning has shifted in a way that would plausibly meet the Functioning Criterion and so bolster a claim to Semantic Understanding. Finally, I consider multi-modal systems such as DALL-E and DALL-E 2 and argue that, with some care, all of the arguments against the Semantic Hypothesis apply to these systems as well. The problems raised for large language models such as ChatGPT are not specific to text-based systems, then, but rather apply in virtue of the statistics-of-occurrence strategy itself.

I conclude by summarizing the key arguments and conclusions of the foregoing, discussing the scope of the argument, and raising some prospects for future research. It is likely that by the time this paper is published there will be even more advanced language models out there. The arguments and lessons of this paper should still largely apply to them. While the ensuing claims may be somewhat pessimistic, the lessons I wish to take from them are not. By getting clearer about what it would take for an AI system to have semantic understanding and for us to detect it, we may hopefully make better progress toward that end.

Before moving on, it is important to clarify terminology. One regrettable aspect of literatures regarding AI generally and language models in particular is that terms evocative of human intelligence such as "attention", "knowledge" and "learning" are often used in quite technical senses that are assumed to have some intimate connection to genuine intelligent attention, knowledge, learning, etc. Since the question of their relationship is precisely what is at issue in this paper, any time one of these terms is used in a technical sense, I will add an "*" to the end of the term. Thus, so called "attention mechanisms" in the literature will be marked as attention* mechanisms, to make salient the concern that the connection between them and genuine attention mechanisms remains an open question.

1. Conditions on semantic understanding

Before diving into the question of whether certain AI models have semantic understanding, we should get some clarity on the question of what semantic understanding involves. I will assume here that humans (and many other animals) have semantic understanding, while many artificial systems such as laptops and toasters, and biological systems such as bacteria and trees, fail to have it. While semantic understanding is a technical term, we can get at our question through commonsense phrases such as: "Does the AI know what it's saying?" "Does the AI understand what I'm saying?" "Is the AI actually talking with me?". Each of these phrases gets at the idea that we are curious about whether the AI system in question is interacting *meaningfully* with its interlocutors, whether the interaction amounts to genuine communication where its behavior is due to an understanding of a human's words and intentions and reflects an effort to respond in kind.

Centrally here, when we ask whether the AI has semantic understanding, at least part of what we are asking is whether its responses

[3] I will not distinguish between various precisifications of this hypothesis in what follows because the criticisms I will develop apply to all such variations. I will discuss the claim that "SOMs have semantic understanding" interchangeably with endorsement of Semantic Hypothesis for ease of exposition.

reflect a *sensitivity* to the meaning of the text it produces. When it writes a poem, or summarizes a piece of text, are its textual outputs a result of understanding the meanings of words and phrases? Or is it merely producing text that looks meaningful?

We can refine this idea as follows:

> **Functioning Criterion**. A system with semantic understanding functions in ways that are causally explainable by appeal to the semantic relationships among its states and processes with semantic content, and this functioning typically drives the evolution of these states and processes as well as the system's overt behavior.

In other words, the Functioning Criterion says that systems with semantic understanding have covert functioning as well as overt behavior that is best explained as being driven by sensitivity to, or respect for, semantic relationships. This is a criterion rather than a definition, for someone might argue that there are further conditions that must be met for a system to have semantic understanding.[4]

However, it is very plausibly a *necessary* condition – a condition that any system possessing semantic understanding must meet.[5] It captures the idea that semantic understanding is not idle; rather, an important aspect of what it is to have semantic understanding is to engage in meaningful reasoning and behavior. It is hard to say that a system's functioning or behavior is meaningful, or reflects understanding of its semantic contents, if these contents do not feature prominently in an explanation of how and why the system behaves as it does.

In what follows, I will explain why SOMs fail to satisfy the Functioning Criterion, even on very generous interpretations of what it takes for the contents of a system's states and processes to be causally explanatory of the system's behavior. In particular, I will allow for our purposes here that systems can satisfy the Functioning Criterion even if, strictly speaking, only internal functional (i.e. formal) properties ultimately drive the system's functioning and behavior.[6] On this interpretation, a system satisfies the Functioning Criterion when the question of *why* the system functions as it does is answered by appeal to the semantic contents of its states and the way its transitions respect semantic relationships among those states. This representational approach is central to the computational turn in cognitive science, and admits of a variety of interpretations of what it is for semantic contents to explain a system's functioning in ways that support claims to that system "respecting" semantic relationships.[7]

While there is substantial debate about exactly how to cash out these ideas, our discussion will not turn on the details. This is because, by focusing on whether a system meets the Functioning Criterion, we can prescind from detailed discussion about what constitutes genuine representation in cognitive systems, and instead focus on a contrastive problem. For our purposes we can count the Functioning Criterion as satisfied if the system in question is best explained by attributing *natural language semantic* properties to its vehicles in providing a causal explanation of its functioning and behavior rather than other kinds of properties. That is, is the system best understood as functioning in ways that track semantic properties? Or, is it better understood as functioning in ways that track non-semantic properties (if it tracks any at all)? Without answering this contrastive question in favor of semantic properties, it is hard to say that a system is sensitive to meaning in the way required for semantic understanding, because its functioning is best explained without reference to these properties. I will argue that this is exactly the position we are in.

2. Why the inference from behavior is not straightforward

In exploring whether SOMs may have semantic understanding, it is useful to begin with discussing predictive Vector Space Models (VSMs).[8] These models, though slightly older and not as impressive as current SOMs, have also received a lot of attention regarding their meaning-semblant behavior. VSMs are a kind of Distributional Semantic Model, and as we will see below the oft-cited "Distributional Hypothesis" underlying Distributional Semantic Models is sometimes invoked to support claims about semantic understanding. Moreover, one of the most careful and detailed arguments that an SOM has semantic understanding was developed for predictive VSMs such as Word2Vec, so investigating these models will help us to better see both the pull of the Semantic Hypothesis and the challenge to it that I wish to advance in this paper.

Predictive VSMs include Word2Vec (CBOW and Skip-gram, Mikolov, Chen et al. (2013) and Mikolov, Sutskever et al. (2013)), Eigenwords (Dhillon et al., 2015), and GloVe (Pennington et al., 2014), among others. These models are trained on large corpuses of natural language data and optimized to predict nearby words in corpus text. The strategy used to accomplish this task is to develop vector representations of words.

These vector representations represent each word in the model's vocabulary as a vector in a multidimensional space. The model is trained to predict the occurrence of nearby words and in the process develops a latent dense vector representation (Mikolov, Chen et al., 2013). Because these models are trained to predict the occurrence of nearby words, the kind of similarity the model is optimized to represent is *word co-occurrence similarity* – roughly, how likely the words are to occur close to one another in text, and how likely the words are to occur in what would otherwise be the same text stream (e.g. "The *policeman* found the suspect" and "The *man* found the suspect").[9] These similarity relationships, then, are primarily similarity relationships of positioning in text. They are not directly semantic similarity relationships — relationships of synonymy or other relationships of meaning.[10]

Generalizing, we can use the term *statistics-of-occurrence properties* for properties such as the frequency of word co-occurrence in text, the similarity of positioning of text in a word corpus, the probability that

[4] For example, one might think that consciousness is required for genuine semantic understanding, or that there are further criteria on understanding the meanings of terms, such as some kind of embodied interaction with the most basic contents of thought. This paper is neutral on these questions.

[5] Semantic understanding is not just linguistic, but it is often manifested linguistically. As cognitive agents, our thought processes and actions reflect an understanding of the world around us, and efforts to interact meaningfully with that world. For example, when my daughter smiles back at me, she is expressing her delight at me smiling at her. Her smiling back at me reflects not only a certain kind of pleasure, but a recognition of me as her parent and my action of smiling at her. Since we are concerned with language models and representations in this paper, we are focusing on semantic understanding as reflected in linguistic behavior. However it is important to acknowledge that the phenomenon is broader, and at root has to do with reasoning and agency.

[6] These *formal properties* are abstractions of physical properties of the system that are mathematically or otherwise functionally specified — 1s and 0s in the most basic case for a binary computer, but typically more abstract properties, such as neural network activation patterns or mathematical entities such as vectors.

[7] See Clark (2001), Ch. 1 for introduction of this idea and description of its importance in cognitive science. See Fodor (1987) for a canonical development of this idea. See e.g. Shea (2018) for a recent theory. For reasons beyond the scope of this paper, I endorse a more robust conception of functional sensitivity to semantic relationships than the representationalist conception. Because of the widespread acceptance of the representationalist conception and its being weaker than the conception I prefer, I adopt it here.

[8] For an overview of early VSMs and their importance in advancing our ability to robustly computationally represent, see Turney and Pantel (2010). For a more recent overview see Lenci et al. (2022).

[9] Similarity measures are typically normalized so that ubiquitous words like "the" and "and" are not considered similar to other words just in virtue of their commonality. See Jurafsky and Martin (2023) for an overview.

[10] It is an interesting empirical question what kinds of semantic information might be carried by the dimensions of a VSM. See Hollis and Westbury (2016) for discussion.

the next word in a string will be a certain word, or the probability that an image or formally definable aspect of an image is correctly associated with a certain caption. These are all formal properties that relate to either frequency statistics or probabilities about text string placement or image-text association. We can accordingly call any model a *Statistics-of-Occurrence Model* (SOM) if it is either designed to represent statistics-of-occurrence information in text or multi-modal corpora, or it is optimized for text prediction on the basis of data about such placement. These are the models of concern in this paper.

Thus, while I focus on predictive Vector Space Models (VSMs) such as Word-2Vec in this section, VSMs as a class count as SOMs in virtue of being designed to represent (frequentist) statistics-of-occurrence properties.[11] As we will see, Language Models (LMs) such as GPT-3 are also key examples of SOMs in virtue of being trained and optimized for text prediction on the basis of data about word placement. In order to keep the discussion streamlined, and to make the case against the Semantic Hypothesis at the level of detail required, I focus largely on text-only SOMs, which have recently generated a lot of discussion regarding semantic understanding and for which we have a more developed literature related to assessing semantic understanding. I briefly discuss multi-modal systems such as DALL-E in Section 5.2 and argue that the concerns presented for text-only systems generalize.

Once a VSM is trained or constructed, researchers can probe the model to see if it carries information about semantic relationships. This is typically done by performing vector algebra with the word representations to try to detect hidden knowledge* within the system (see e.g. Mikolov, Yih et al. (2013)). So, for example, in order to get the model to solve* for an analogy such as *Switzerland is to Swiss as Cambodia is to [BLANK]*, one might compute:

$$vec(Switzerland) - vec(Swiss) = vec(Cambodia) - x$$

where the vector for “Switzerland” is vec(Switzerland), etc. Then one solves for x among vector representations of the model. If the closest word vector is for “Cambodia” is “Cambodian” (typically measured by Cosine similarity), then the analogy is considered to have been correctly solved.[12] The thinking behind these kinds of tasks is that the distance between vec(Switzerland) and vec(Swiss) should be close to the distance between vec(Cambodia) and vec(Cambodian), and indeed that distance should be more similar than any from vec(Cambodia) to any other word in the vocabulary.

Authors report impressive results. Language models can answer* both syntactic analogy questions, like the nationality question above, as well as more straightforwardly semantic questions, such as *brother is to sister as grandson is to [BLANK]*. For example, Mikolov, Chen et al. (2013) report that CBOW had an accuracy of 24% on the semantic test set Mikolov et al. created, and 64% on the syntactic test set, while Skip-gram had an accuracy of 55% on the semantic test and 59% on the syntactic test set. These results are interesting in themselves, simply because they show that language models can carry both syntactic and semantic information when optimized to do something else, namely predict nearby words.[13]

It is important, however, to distinguish the sense in which VSMs can be said unequivocally to *carry information* from claims to *semantic representation* which might be taken to constitute, or help to underlie, semantic understanding. As I will use the term in this paper, language models *carry semantic information* just in case formal properties and relationships internal to the system are sufficiently correlated with semantic properties and relationships. So, in this case, we can say that CBOW carries information about countries and nationalities because vector algebra that exploits formal relationships between word vectors reveals systematic correlations between these formal properties and semantic properties such as country-nationality relationships. We can say in such cases that the formal features internal to the system (sub-states and processes), whose relationships correlate with semantic relationships, carry that semantic information. So, the vector for “Cambodia” carries information about the country Cambodia, and the vector for “Cambodian” carries information about the nationality.

Carrying information, in this sense, is an important property, but does not itself entail features that are often thought to be important for a feature of a system to count as a representation of the sort that would vindicate claims to semantic understanding. For example, it does not entail that internal representations are appropriately causally connected to the features they purportedly represent (Fodor, 1987; Neander, 2017; Stampe, 1977). In this paper, I focus on the gap between a vehicle’s carrying information and its *functioning* within the system to carry semantic information (see Dretske (1986), Millikan (1989) and Shea (2018)). Crucially,

just because there are internal formal features that carry information does not mean that the system is causally well-described in terms of variables that represent that information, which can be manipulated in order to alter the functioning of the system (Geiger et al., 2021). So, while we can and should hold that VSMs carry semantic information in the sense just specified, that does not yet settle the question of whether SOMs satisfy the Functioning Criterion. We need to investigate whether systems of this kind plausibly function in ways that are causally explainable by appeal to their semantic properties.

Focus on the Functioning Criterion helps to bring into relief that evidence of meaning-semblant behavior does not straightforwardly support that a system satisfies the Functioning Criterion. A system can exhibit meaning-semblant behavior and still lack semantic understanding, so we must investigate whether the best explanation of this behavior, *given what else we know about the system in question*, involves attributing semantics-sensitive functioning to it.

Acknowledging a gap between meaning-semblant behavior and semantic understanding follows from general claims about complicated or complex systems. There are multiple possible functional structures that can instantiate any but the most simple stimulus-behavior pattern. As such, context and other evidence about system functioning is crucial for understanding system functioning and so for attributing semantic understanding. Appreciation of this fact for semantic understanding was one of the chief lessons of criticisms of behaviorism. It is generally impossible to attribute beliefs and other meaningful psychological states to systems just on the basis of behavior alone, for our beliefs interact in complex ways with our hopes, desires, background beliefs, etc. (Chomsky, 1959).

We can acknowledge that there is a gap between meaning-semblant behavior and semantic understanding without embracing skepticism about other minds, animal cognition, or genuine artificial intelligence.[14]

[11] Older VSMs that do not use a predictive training strategy, but rather develop vector representations by counting occurrences of words, therefore also fall within the scope of the argument. Such models include Hyperspace Analog of Language (HSA, Lund and Burgess (1996)) and Latent Semantic Analysis (LSA, Landauer and Dumais (1997)). See Gunther et al. (2019) and Lenci et al. (2022) for overviews. Thanks to an anonymous referee for encouraging clarification on this point and greater precision in my discussion of VSMs and LMs.

[12] See Jurafsky and Martin (2023), Ch. 6. See Linzen (2016) and Rogers et al. (2017) for critical assessment of this evaluation strategy. In particular, where answers are very similar to one another such as “Switzerland” and “Swiss” the method of finding the closest word vector to prescribed point by cosine similarity may just be the closest neighbor to the target word (in this case “Cambodia”) and not reflect a sameness in structural relationships between the two.

[13] While for the purposes of this paper I grant impressive meaning-semblant behavior of these models, the extent of this performance and the best means of assessing it are subjects of debate. See Church (2016), Linzen (2016) and Rogers et al. (2017) for some representative criticisms.

[14] Thanks to an anonymous referee for suggesting I discuss this question.

We do not, outside of the philosophy classroom, seriously question whether we have enough evidence about other people to attribute mental qualities such as consciousness and semantic understanding to them, and we regularly attribute these qualities to animals. Plausibly, however, this has to do not just with a lowering of the evidential bar from these contexts (where absolute certainty may be required) but also with (1) a huge swath of behavioral evidence, (2) certain claims to internal functionality being *already* supported by our own introspective case, and (3) evidence of similarity in neural and biological functioning among humans. Despite our diversity and difference, humans are quite similar to one another in terms of both behavioral patterns and covert neural and cognitive functioning. For only one human (namely the subject of the skeptical question-asking) to be the only one to have a mind is not the simplest or most parsimonious explanation. Instead, the subject has no reason to think she is special. The best explanation is overwhelmingly to attribute mentality to other humans.

This reasoning extends, with some nuance, to animals. Animal ethology has gradually seen acknowledgment that the best explanation of animal behavior typically involves attributing semantic understanding and in many cases complex cognition to animals. Simpler explanatory paradigms will not do the trick (Graham, 2019). Psychological paradigms where the Functioning Criterion is met seem to be the most parsimonious and cohere best with our other scientific and world knowledge. Moreover, we have evolutionary and neuroscientific evidence that the animals are built similarly to us, which supports claims that such systems psychologically function similarly (see Andrews (2020), Ch. 3 for an overview). To the extent that this is attenuated (e.g. octopi), stronger behavioral or other evidence is often required for the attribution of cognition or semantic understanding (see Godrey-Smith (2016) for discussion).

In the case of artificial systems we take the same empirical and non-skeptical approach; however, the evidential situation is different. Our behavioral evidence is substantially restricted in comparison with other humans and animals (in the case of SOMs typically to verbal or pictorial interactions through a user interface), and we cannot assume similarity in internal functioning due to shared evolutionary history and pressures. Moreover, our tendencies to anthropomorphize may lead us astray.

Work on anthropomorphism of AI systems over the last 50 years strongly suggests that we have a tendency to over-attribute agency and understanding to inanimate and unintelligent systems when they merely exhibit some meaning-semblant behavior. Block (1981), for example, importantly remarked on this tendency regarding the AI program *Eliza* built by Weizenbaum (1966), a program which few today, if any, would regard as having genuine semantic understanding, but which elicited very strong intuitive reactions from interacting subjects at the time. It is widely believed that attributions of agency and explanations from the “intentional stance” (Dennett, 1987) are a default and basic psychological process, helping to explain attributions of animacy to natural entities (trees, mountains) and events (rain, volcanoes) (Mitchell et al., 1997). Additionally, we know that features that are not essentially indicative of intelligence can drive our judgments about agency and semantic understanding. Morewedge et al. (2007), for example, showed that behavior on a fast timescale increases attributions of mind to robots. Epley et al. (2007) argue that our tendency to anthropomorphize is increased by our own desire for social contact and affiliation, which is independent of the capacities of the AI system.

This all suggests that the *interpretability* of AI behavior as meaningful is motivated by a variety of features which do not necessarily indicate that the behavior is indeed meaningful, but this interpretability itself strongly motivates attributions of mentality, agency and understanding. So, although merely behavioral criteria for intelligence have a distinguished history (Turing, 1950), given what we now know about proper attributions of mentality to humans, animals, and artificial systems, we should acknowledge that meaning-semblant behavior can come apart from genuinely meaningful cognition and behavior and approach the question of semantic understanding in a way that takes into account our best understanding of system functioning.

The foregoing are general concerns about over-attributing semantic understanding to AI systems, which serve as important context; however, our focus here is on SOMs in particular. In their case, we have special reason to doubt the inference from behavior to internal functioning, and so semantic understanding.[15] This is because we plausibly have a better explanation of SOMs’ meaning-semblant behavior. Because humans use language meaningfully and systematically, there are strong correlations between word co-occurrence statistics and semantic relationships. A model trained to represent statistical co-occurrence information or optimized for word prediction over a large text corpus will thus tend to carry semantic information merely for these reasons. Thus even if its functioning is only driven by sensitivity to word-co-occurrence statistics, it will still demonstrate meaning-semblant behavior.

Because these facts are widely agreed to, even by proponents of the Semantic Hypothesis, if it is plausible that merely attributing sensitivity to word co-occurrence statistics is sufficient to explain SOMs’ production of meaning-semblant behavior, then such an account is preferable, because it is simpler and more parsimonious. I call this alternative explanation the *Statistical Hypothesis*.[16]

3. Why exhibited meaning-semblant behavior is not enough

So far, I have argued that we should not straightforwardly infer semantic understanding from the meaning-semblant behavior of SOMs. Instead, we should aim to make an inference to the best explanation based on what we know about semantic understanding (e.g. that it requires satisfying the Functioning Criterion), the system’s behavior, and what we can plausibly infer about system functioning. A natural response at this point would be to maintain that the kind of meaning-semblant behavior SOMs exhibit is indeed more plausibly explained by a revision of our understanding of their internal functioning, so that it satisfies the Functioning Criterion. This response disputes the claim that the Statistical Hypothesis can better explain the exhibited meaning-semblant behavior.

It is impossible in a paper-length treatment, let alone one that aims to canvass other options for defending the Semantic Hypothesis, to examine all the kinds of behavior that one might claim cannot be explained by appeal to functioning that tracks statistics-of-occurrence properties rather than semantic properties. Instead, I will provide two illustrative examples that I hope will make it plausible that SOM behavior can be so explained. Given the straightforward availability of the Statistical Hypothesis, I hope that establishing the plausibility of this claim will put the ball back in the court of the proponent of the Semantic Hypothesis. Perhaps there is some meaning-semblant behavior of SOMs that directly support claims to its meeting the Functioning Criterion, but making this case is harder than one might think.

Let us then, explore our first example. Bhatia (2017) has provided, to my knowledge, the most developed and rigorous behavioral defense of an SOM’s claim to semantic understanding. He argues that the way VSMs like Word2Vec exhibit certain patterns of behavior supports the claim that they underlie our capacities for analogical, and more

[15] Thanks to an anonymous referee for pushing me to clarify these points.

[16] I am not the first person to claim that SOMs fail to have semantic understanding because the way they produce their responses fails to be sensitive to meaning. See Bender et al. (2021), for example, who assess language models as mere *stochastic parrots* given our understanding of how they are trained and how they function. I also ignore in this paper the many behavioral shortcomings of SOMs which the Statistical Hypothesis is also better poised to explain. See Church (2016) for discussion of some behavioral shortcomings of VSMs and Merrill et al. (2021) for discussion of some behavioral shortcomings of transformer-based SOMs like GPT-3.

generally, heuristic, reasoning. He showed that the profile of behavior of VSMs is similar to the profile of behavior of humans performing heuristic and other kinds of reasoning.[17]

Bhatia presented* VSMs with the famous and well-replicated "Linda" problem of Tversky and Kahneman (1983), which is widely taken to reveal important features of human capacities for heuristic reasoning. Bhatia showed that VSMs produce similar results to humans and suggested on this basis that human heuristic judgments may be explained by underlying vector representations within human cognitive systems.

In the original experiment, human subjects are presented with a description of a woman named Linda:

> Linda is thirty-one years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations (p. 297).

People tend to judge that it is more likely that Linda is a feminist bank teller than that she is a bank teller. That is, however, impossible. Every feminist bank teller is a bank teller, and so it cannot be more likely that Linda is a feminist bank teller than that she is a bank teller. Once the facts about probabilities are laid out, this is a simple logical inference. However, people reliably give the impossible answer. Why do they do this? On what basis are they making this inference?

While there is little agreement on the exact solution, many researchers think that our answer to the Linda problem has to do with taking *being a feminist bank teller* to be more representative of Linda as described than *being a bank teller*. The heuristic people use to solve the problem has thus been called the *representativeness heuristic*. What algorithms or processes might underlie this heuristic? It has long been suggested that a more connectionist, or associative, process, may be at play for the representativeness heuristic, while a more logical or symbolic process may be at play in finding the correct solution (Barbey & Sloman, 2007; Sloman, 1996). Perhaps, then, language models of this kind could subserve heuristic reasoning such is at work in the Linda case.

Bhatia shows that three different VSMs (Word2Vec, Eigenwords, and GloVe) give the same answers* to the Linda Bank Teller problem that humans do. He uses vector averaging to come up with vectors for the description of Linda, and he does the same for "bank teller" and "feminist bank teller". The answer vector that is closest (by cosine similarity) to the description vector is taken to be the VSM's answer. So, if the vector for "feminist bank teller" is closer to the description vector than that for "bank teller", the VSM is taken to have answered* similarly to humans. And this is exactly what he finds for all three VSMs studied.

Additionally, he finds that VSMs exhibit a lot of other behavior similar to that of human subjects. For example, they do not find a closer association between the Linda description and "feminist bank teller" than they do between it and "feminist". Furthermore, VSMs provide similar responses to similarly constructed problems. They tend to give answers* compatible with base rate neglect similarly to humans. Impressively, they perform well above chance on real-world question and answer data sets, and correct performance by the VSM is predictive of whether humans on average get the correct answer. This is important because it is widely hypothesized that our judgments in the Linda case are a result of more general reasoning processes that are reliable and adaptive (see Kahneman (2011) for discussion).

This line of argument is importantly more detailed than standard academic arguments in favor of the Semantic Hypothesis, which tend to focus on SOMs exceeding a certain level of performance on benchmark tests (such as TriviaQ&A, Brown et al. (2020)). Bhatia has done something much more nuanced here, which is to show that VSMs' *profile* of behavior is similar to humans' profile of behavior, in terms of both correct and incorrect answers. For this reason, one might be especially moved by Bhatia's case that VSM architecture subserves heuristic judgment in humans.

However, this behavior is well-explained by the Statistical Hypothesis. Characterizations like Linda's are more likely to co-occur with "feminist" than with "bank teller" or "feminist bank teller". Thus a system that is merely sensitive to statistical co-occurrence properties could provide the same pattern of behavior exhibited without functioning sensitively to semantic properties, such as judging the description to be more representative of being a feminist than a bank teller. The same goes for similarly constructed problems. We do not need to explain the functioning of SOMs by appeal to anything like a sensitivity to a representation of what it is to be a feminist. Quite impressively, word co-occurrence relationships are enough. Thus the kind of evidence Bhatia adduces does not give us special reason to think that the Functioning Criterion has been met. The Statistical Hypothesis can still straightforwardly explain this behavior.

Similar considerations apply to the kind of meaning-semblant behavior originally adduced in favor of the Semantic Hypothesis for VSMs described above, namely their probe-ability for semantic information using vector algebra. While this behavior is indeed impressive, it can plausibly be explained without the system functioning sensitively to semantic properties per se. Indeed, despite the excitement about VSMs as semantic models, this is plausibly the straightforward interpretation of the Distributional Hypothesis (Firth, 1957; Harris, 1954; Sahlgren, 2008) and its motivation of VSMs as representations of word meaning in Machine Learning, NLP, and AI research. According to the Distributional Hypothesis, words with similar positions in text corpora tend to have similar meanings, and so representations of this information can serve as proxies for word meaning. This is a hypothesis about a *correlation* between positions in text corpora and word meaning that researchers can exploit, because neural networks can be trained more directly to track these statistical properties. So, the vector algebra that mirrors analogical reasoning does so by exploiting abstract similarities between the relationships in relative positioning between two pairs of words, for example "Switzerland"/ "Swiss" and "Cambodia"/ "Cambodian". This kind of similarity correlates highly with nationality relationship but representing nationality per se is not a property that needs to be invoked in explaining how the system functions to solve* the analogy. The word co-occurrence properties do all the work.

The case that VSMs do function sensitively to semantic properties *per se* has not, to my knowledge, been clearly argued for, though I

[17] Another strategy based on behavioral evidence looks for tests or markers of a particular aspect of our semantic capabilities, such as the capacity for compositionality in thought and language, and tries to assess whether AI systems pass these tests (see e.g. Baroni (2019) and Hupkes et al. (2020)). While this approach is important and generates many valuable insights, I take a different approach here which I think is under-explored and has its own substantial benefits. By keeping the discussion focused on general claims to semantic understanding and the more general issue of semantic functioning, we can avoid debates about the relative importance or uniqueness of these specific capacities for semantic understanding, or the convincingness of various tests. Moreover, and perhaps more interestingly, we can ask and answer questions about semantic understanding in a way that does justice to what many perceive to be so powerful about these systems: that their size, and the size of their training data, enable SOMs to transcend the limitations of specialized systems and exhibit domain-general capacities in a way that might have special claim to semantic understanding. Thus a focus on general claims about the behavior of these systems, and what their implications might be for their internal functioning, are worth considering. In light of these interests, Bhatia's work is especially interesting and pertinent, because the capacity for analogical, or heuristic, reasoning does not have a specific content or form, and his motivations for attributing this capacity to SOMs do not rely on specific grammar or content capabilities but rather with the kinds of general capabilities we expect VSMs to have in virtue of their training and functioning. Thanks to an anonymous referee for urging me to address this alternate approach to evidence for semantic understanding.

will explore some complexities below. Once we lay out the Functioning Criterion as a plausible criterion of semantic understanding, and articulate the Statistical Hypothesis as a distinct hypothesis, it seems superfluous to advance the Semantic Hypothesis on the basis of this kind of behavioral data.

3.1. GPT-3 to the rescue?

Perhaps, then, an example of more complex and impressive behavior such as that of GPT-3 will pose more of a challenge to the Statistical Hypothesis. SOMs like GPT-3 use *autoregressive generation* to produce strings of text, where they produce the next word or sequence based on an assessment* of the highest probability word based on the previous word sequence(s) (see Jurafsky and Martin (2023), Ch. 9 for an overview). Current machine translation, summarization*, and question answering* systems all use autoregressive generation.

Where advanced VSMs like Word2Vec are optimized to predict nearby words, and in the process develop latent vector representations of words, autoregressive language models like GPT-3 assign probabilities to words in the vocabulary based on preceding text. They use a transformer architecture to dynamically assess* the probabilities of words, and so to provide appropriate responses to questions based in part on examples (one- or few-shot learning). The details of the architecture are unimportant for our purposes, except that they enable the language model to autoregressively generate text based on the differentially weighted influence of previous text.

While transformer-based architecture is significantly more sophisticated than predictive VSM architectures like Word2Vec, the most straightforward interpretation of these systems is that they function in ways that are sensitive to the statistics-of-occurrence properties of text, since those are the ones that are directly relevant to word prediction on the basis of text string data. The transformer architecture plausibly functions to encode more fine-grained and contextualized information of the same kind, not directly semantic information.[18]

Even important capacities such as text summarization*, that might seem to train specifically for or require semantic understanding, do not plausibly do so upon further investigation. Language models that perform text summarization are trained on pairs of full length articles and their summaries. They are still optimized for prediction (in this case of the summary text based on the full text) and can be made to produce summaries using autoregressive generation in the same way as other text generation.

One might object — but how could GPT-3 and similar models successfully perform text summarization without having developed functional sensitivity to semantic properties? After all, the capacity to summarize seems to involve being able to convey the meaning of a longer text in a shorter one, and so it seems like the model must *somehow* have developed sensitivity to semantic relationships.[19] However, it is not at all obvious that GPT-3 actually has this semantic capacity, *per se*. Texts tend to include signposts or other cues about main ideas that an SOM can pick up on, from the number of times a term is mentioned, to explicit statements of what the text is about that might involve markers like "I will argue that" or "We'll see that", to first sentences of paragraphs including more pertinent and general information. An SOM can pick up on all this correlated statistics-of-occurrence information.

Moreover, it is likely that such texts can represent statistics-of-occurrence-based similarity relating shorter and longer text strings. To the extent that a shorter text is similar to a longer text in the sense of tending to occur in similar contexts, this might be a good proxy for sameness of meaning, hence summarization. It is also often hard to tell whether a text is entirely novel to the system, or whether it bears close similarities with training samples. For more common texts and summarization tasks, some of the success may be due to training on very similar texts. (This is increasingly so as users interact with the system, generating more data.) Finally, and perhaps most importantly, the quality of many summaries is not great, especially for topics that are rare and thus unlikely to be well-represented in the training data (such as obscure philosophy papers).[20]

So, while GPT-3 is behaviorally quite impressive – it can perform well on trivia (TriviaQA), unscramble words, do* arithmetic, successfully use* novel words in a sentence after only one example, and generate synthetic – read *fake* – news articles that are difficult to distinguish from human-produced ones –[21] these behavioral capacities do not challenge the basic point. Given the strong correlations between statistical co-occurrence relationships and semantic relationships, a system optimized for predictive accuracy can reliably produce behavioral outputs that *look* meaningful without actually functioning in ways that are sensitive to semantic information.

So, while these considerations are of course not comprehensive, I hope they establish the Statistical Hypothesis as plausible enough that the proponent of the Semantic Hypothesis must adduce *additional* evidence than the kind of behavioral evidence so far presented. We should look to more direct arguments about SOMs' functioning that supports their meeting the Functioning Criterion.

4. Are semantic properties really so distinct?

The second response in defense of the Semantic Hypothesis claims that functioning sensitively to statistics-of-occurrence properties is actually sufficient for satisfying the Functioning Criterion. *All it is* to have semantic understanding on this view (at least for a range of capacities of interest) is to have an architecture that is optimized for statistical accuracy. This idea is worth taking seriously in large part because it is sometimes invoked by appealing to the Distributional Hypothesis, described above.

Although the Distributional Hypothesis is a correlational thesis, one might be interested in a stronger reading of it on which *all it is* for a word to have meaning is for it to have a certain global pattern of use in a language, so that a representation of statistics-of-occurrence information is *thereby* a representation of word meaning. While there

[18] It is an interesting question whether the incorporation of other strategies, such as so-called active learning* (e.g. along the lines of Mnih et al. (2014)), could substantially change the architecture of a language model so that its functioning more plausibly lays claim to semantic sensitivity. While this is intriguing, I doubt there will be any quick fixes here. For example, as long as the models are optimized for text prediction and have as data information about text placement, functioning that preferentially selects certain inputs in the learning process will not thereby lead to semantic sensitivity. That is not to say that there cannot be other uses for active learning that might help a system to meet the Functioning Criterion, but discussion of these possibilities is beyond the scope of this paper.

[19] Thanks to an anonymous referee for encouraging me to discuss this question.

[20] Very unscientifically the author asked ChatGPT to summarize a draft section of this paper. Here is an excerpt, "Furthermore, the text highlights the need to establish that vector representations built on statistical co-occurrence information represent the standard referents of words and not just carry information about them. It suggests that proponents of semantic understanding in Statistical Outcome Models (SOMs) should aim to show that these models possess the features required for genuine semantic understanding". This sort of looks like a summary of part of this section, until one looks more closely. The term "representation" is focused on but is not particularly relevant. (We are not concerned with representation per se but with the functional capacities of systems.) It also uses the term "Statistical Outcome Models" which is used nowhere in the paper. And it states generally that proponents of Semantic Understanding should aim to show that these models possess the features required for genuine semantic understanding which is true but much less specific than the claim of that section, which is about satisfaction of the Functioning Criterion.

[21] Brown et al. (2020).

may be some proponents of this claim, it is highly contentious, in large part because it ignores the kinds of external relations to the world that theorists tend to think are required for genuine content. For example, in order to have a representation that means *dog*, it is plausible that the representation must bear some relation to dogs (see Gasparri and Marconi (2019) sect. 3.3 for discussion.) This is not entailed by statistical co-occurrence information in text corpora.[22]

Instead, I think that a more nuanced and interesting idea is often behind the stronger reading of the Distributional Hypothesis, namely that meaning *at least partially* depends on statistics-of-occurrence information, and that this is enough to make it the case that a model that represents (or is trained to represent) the latter can thereby represent the former in the robust sense required for semantic understanding (see Lenci (2008) for discussion). So, interpreted as a defense that SOMs satisfy the Functioning Criterion, the claim would be that a system that functions sensitively to statistics-of-occurrence properties would thereby function sensitively to meaning. This is an interesting hypothesis, especially given the close correlations between these two kinds of properties.

The problem with this response is that it, despite the close correlations, there are still important differences between functional sensitivity to one rather than the other. To see the distinction at issue here and how it can get elided, consider some supplementary motivation Bhatia provides for his view. Bhatia claims that work in cognitive science supports his view, in particular Shafir et al. (1990)'s claim that what drives the results in cases like Linda's are *typicality judgments*. According to Shafir et al. when one reads the Linda description and available answers (e.g. feminist, bank teller, feminist bank teller), one judges (unconsciously) the degree to which the answer is typical of a person satisfying the description, and chooses the answer with the highest degree of typicality. Bhatia claims that his results are coherent with this work.

However, there is an important difference between Shafir et al.'s hypothesis and Bhatia's. Typicality judgments are importantly meaningful – i.e. semantic. When we make typicality judgments, we are judging, *of people* described in a certain way, whether it is typical that they would have the property in question (feminist, bank teller, etc.). Typicality judgments inherently involve *predication* – taking Linda to *have* the property of being a feminist because she *has* the properties attributed to her in the description. Functional sensitivity to predicative information – i.e. to *being* a feminist rather than a bank teller – requires sensitivity to the conditions under which one would be a feminist or a bank teller, which goes beyond sensitivity to statistics of text corpora and into the world.

Moreover, one should not ignore the semantic understanding that goes into *identifying* Linda, feminists, or bank tellers *as such*. This ability to identify people or the instantiation of properties in the world is required even where one is processing experiential co-occurrence statistical information. However, it is not required for processing co-occurrence information in text corpora. To be functionally sensitive to, for example, the co-occurrence of tables and chairs, requires functional sensitivity to tables and chairs themselves, not just the words "table" and "chair" (or images of them). It pertains directly to *referents*, not words or phrases. Our typicality judgments generally involve this more basic semantic understanding as well.

Proponents of this second response, by claiming the sufficiency of functional sensitivity to statistics-of-occurrence properties for functional sensitivity to semantic properties, in fact deny that we have much of the functional sensitivity that is required for genuine semantic understanding. They therefore, perhaps accidentally, endorse an *error theory* of our own semantic understanding. By claiming that our behavior is in fact driven only by functional sensitivity to statistics-of-occurrence properties in these cases, they are committed to the view that what seems like actual sensitivity to the referents of our words in the world is merely a much more proximal kind of sensitivity.

I take it that one should avoid error theories of our semantic capacities if one can. Moreover, dialectically the position is precarious. The big, interesting claim that we were meant to evaluate is that AI systems can meaningfully interact with their environments just like humans and many animals do (albeit in a restricted way, through text exchange), not that humans and many animals in fact fail to meaningfully interact with their environments in the way we normally attribute to them. When we water down our conception of what semantic understanding requires, we make it more likely that AIs can meet our criteria, but we make a much less interesting and exciting claim.

Perhaps for this reason, most researchers who support the Semantic Hypothesis do not take themselves to endorse a revisionary account or error-theory of our semantic understanding. They do not, in general, present themselves in this way. Instead they present themselves as claiming something stronger and much more interesting, namely that SOMs have understanding in the robust sense normally attributed to humans. Let us, then, turn to the strongest strategy for attributing semantic understanding to SOMs.

5. Might optimizing for predictive accuracy have created systems with semantic functioning?

The third, and most interesting, response to the challenges I have so far presented is to argue that the *way* advanced SOMs are so effectively predictive is because they learn to pick up on semantic relationships, and they use these in producing behavior. This response claims that the development of functional sensitivity to semantic relationships is *a consequence* of optimizing for prediction. If this is true, then perhaps the Functioning Criterion is satisfied, and we might be justified in attributing semantic understanding, or at least a proto-version of semantic understanding to SOMs.

This is an interesting possibility, and one that deserves careful attention. What reason do we have to think that SOMs function in ways that are genuinely sensitive to semantic content? Chris Potts, in a recent talk, for example, suggests that recent work on Causal Abstraction Analysis, which aims to provide high-level functional analyses of deep neural net systems, might provide us with the kind of justification we need to vindicate claims to semantic understanding for SOMs like BERT and GPT-3 (Potts, 2022).

Causal Abstraction Analysis (CAA) aims to accurately model the high-level functioning of a deep neural net (DNN). According to this approach, (1) one first specifies a candidate causal model that might accurately explain the functioning of the neural network system. Then (2) one identifies potential structures within the neural network that correspond to the high-level variables specified by one's model. Third, (3) one performs "interventions" on the neural network system to test whether changing certain variables has the effect predicted by one's model.

So, for example, one might (1) hypothesize that a DNN performs addition and create a model of the relevant high-level variables responsible for the output. The next step (2) would be to search for corresponding structures internal to the model. Step (3) would be to perform interventions on those corresponding neural net structures to see if they behave as the model predicts.

Potts suggests that, if CAA produced a model of an SOM on which the SOM has representations with the kinds of contents the language

[22] Wittgenstein (1953) is sometimes cited in defense of this stronger interpretation of the Distributional Hypothesis, but his conception of use was purposefully much more expansive. For Wittgenstein, the interpersonal social activities we engage in when we use words are of the utmost importance, as is the diversity of kinds of interaction. He resisted any unitary formal operationalization of word meaning. See Biletzki and Matar (2021), also sect. 3.3, for discussion. We must be careful, then, not to over-state the case for this stronger view.

model is supposed to represent, and it describes a causal structure within the SOM on which it is prone to processes involving these representations that respect their semantic properties, then such a model could make claims to having semantic understanding. As we have framed the issues here, such an analysis might help to justify the claim that an SOM satisfies the Functioning Criterion because it would identify semantic variables that are causally relevant to the SOM's functioning.

To my knowledge, no one has attempted to construct any CAA models for GPT-3 or its descendants, such as ChatGPT. However, Geiger et al. (2021) make some relevant claims for BERT (Bidirectional Encoder Representations from Transformers, Devlin et al. (2018)). BERT also uses transformer-based architecture, but it is *bi-directional*, so that its transformers can make use not just of past information in a text string, but subsequent information as well. The model learns to predict masked tokens (such as "The cat [BLANK] on the mat") and also subsequent sentences. BERT is thus clearly an SOM, and so if Geiger et al. (2021)'s work supports the claim that BERT can satisfy the Functioning Criterion, that would be important evidence in favor of the Semantic Hypothesis.

Let us then see what evidence this work on BERT might provide for a CAA justification of semantic understanding in SOMs generally, including GPT-3. Geiger et al. (2021) evaluate a version of BERT that is fine-tuned on the Multiply Quantified NLI datasets which consist of templated English language examples, labeled according to their linguistic tree structures. They develop a CAA as described above to partially explain some of the results of this fine-tuned BERT. They compare it to some other SOMs for which they cannot develop nearly as robust a causal model.

While the kind of strategy Geiger et al. advocate is promising, there are reasons to be concerned about whether (i) the strategy achieves strong justification for an SOM satisfying the Functioning Criterion, and (ii) whether the success they describe with this fine-tuned version of BERT supports the more general claim that it is *in virtue of* being trained to optimize statistics-of-occurrence properties that it picks up on and functions sensitively to semantic properties.[23]

First, we must be careful in selecting *which* hypotheses to evaluate. Even if such a model were to make accurate predictions about the behavior an SOM under interventions, we would still need to rule out other models that do not attribute inference-like behavior and may be better fits. The important alternative to rule out, in this context, is that the most accurate causal abstraction model is one whose causal variables primarily reflect processes driven by word co-occurrence statistics rather than semantic relationships – i.e. a model that would support the Statistical Hypothesis. Because these properties are so closely correlated with each other, in order to support the Semantic Hypothesis, a CAA analysis must show that we *better understand* the functioning of the system by attributions of semantic content-respecting inferential roles than word co-occurrence-regularity respecting causal roles. It is not enough, then, to develop a causal model in line with the semantic properties one hopes the system represents. One must show that this model better explains the functioning of the system than one in line with a motivated, more parsimonious interpretation.

The second concern, about the success of the case (Geiger et al., 2021) describe extending to other SOMs, is important because the version of BERT they assess was fine-tuned on hand-labeled data *precisely in order to train it to detect certain semantic structures*. One would expect that the functioning of the resultant system would pick up on some of the features relevant for labeling these templated English sentences in accordance with their linguistic trees. In contrast, VSMs and GPT-3 are not fine-tuned in this way, and it would be impossible and defeat the point to try to fine-tune an SOM on all semantically relevant properties of interest. The interesting claim that we have been considering is whether, *merely in virtue of being optimized for predictive accuracy*, these SOMs develop internal functioning that validates claims to semantic understanding. This is because claims that large language models have, or are on their way to possessing, semantic understanding, are in part based on their generality — their ability to chat and answer questions about a wide variety of topics in ways that look meaningful. Claims that they satisfy the Functioning Criterion would then have to come from general features about their design and training, not fine-tuning for specific semantic properties. Geiger et al.'s results do not directly support such claims, as the BERT system they analyzed was fine-tuned with supervised learning precisely to make it function to detect a particular kind of semantic structure.[24]

To summarize, while I have not shown that SOMs definitively lack the features required for semantic understanding, current research does not give us strong reason to be optimistic. We do not have evidence that optimizing for predictive accuracy in SOMs thereby trains systems to function in ways that respect the semantic properties of the natural language words they are trained on. In the absence of such justification, we should continue to accept the Statistical Hypothesis.

5.1. But what about ChatGPT?

On November 30, 2022, OpenAI released ChatGPT, a chatbot AI based on GPT-3.5, a development of GPT-3. The ability of ChatGPT to produce extended, meaning-semblant text, on an expansive array of topics, has shot the question of whether SOMs have semantic understanding into the public sphere in a way that it simply was not before the release. Should we consider the Statistical Hypothesis to still be the default hypothesis even given the behavior demonstrated by ChatGPT?

The argument of this paper has not depended on specific behavioral capabilities of SOMs, but rather on the existence of an alternative hypothesis for meaning-semblant behavior. So if ChatGPT poses a distinctive challenge, it will need to be made on the basis of the way ChatGPT was trained, or analysis of how it functions. As of the time of submission of this paper, there has been no published causal analysis of the internal functioning of ChatGPT. OpenAI has not published a research paper on ChatGPT, but did publish a blog post describing how ChatGPT was built (OpenAI, 2022). Researchers began with GPT-3.5, a development of GPT-3, and then fine-tuned it using Reinforcement Learning from Human Feedback (RLHF). Human trainers provided model examples of human-AI chatbot conversations as data for supervised learning by ChatGPT. They also provided rankings of possible responses to model human questions. This was used to train a reward model, which ChatGPT was then trained to optimize. OpenAI describes this process as similar to the one used for InstructGPT, which does have a corresponding research paper (Ouyang et al., 2022).

Could this fine-tuning be sufficient to shift the internal functioning of ChatGPT so that it plausibly has internal features that function as semantic representations? If a good case can be made that the kinds of methods used on InstructGPT modified its functioning so that it is sensitive to semantic properties, then perhaps we should consider InstructGPT and ChatGPT to be candidates for having Semantic Understanding.

How might such a case be made? Human trainers on InstructGPT were instructed to prefer responses that were truthful and harmless, and secondarily to prefer helpful responses. Could this fine-tuning suffice for adapting the functioning of the system so that it is sensitive to semantic properties? After all, in order for a system to respond *truthfully*

[23] There are many specific questions we can ask here about whether the analysis they provide of this fine-tuned version of BERT is sufficient for semantic understanding — how comprehensive of a model do they need, for example? However, our concerns are much more general, and so we can prescind from this more specialized discussion.

[24] So, does this fine-tuned BERT have understanding? Plausibly not, as it is too specialized. But arguing for that is outside of the scope of this paper.

as such, one plausibly has to act sensitively to the meanings of queries and answers.[25]

At first, things look somewhat promising, as Ouyang et al. (2022) report a *doubling* of truthfulness of InstructGPT over GPT-3 (against the TruthfulQA benchmark Lin et al. (2021)). However, nearly all of this increase is due to InstructGPT producing uninformative answers such as "I have no comment" when it would otherwise have provided an incorrect answer (p. 13). This strongly suggests that fine-tuning based on the truthfulness criterion, instead of significantly shifting the functioning of InstructGPT to make it more inference-like, plays a more isolable role that may be characterizable in non-semantic terms, such as preventing outputs are not highly predicted by the input.

This assessment is also consistent with Open AI's own description of ChatGPT's behavior:

> ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers. Fixing this issue is challenging, as: (1) during RL training, there is currently no source of truth; (2) training the model to be more cautious causes it to decline questions that it can answer correctly; and (3) supervised training misleads the model because the ideal answer depends on what the model knows, rather than what the human demonstrator knows.

OpenAI's admission that there is no source of truth for ChatGPT reinforces our assessment that the fine-tuning ChatGPT receives does not significantly change the internal functioning of the system in the way that would be required for it to actually be sensitive to the meanings of natural language.

While the fine-tuning ChatGPT received makes it produce more plausible-sounding responses and prevents it from producing some harmful and untruthful responses, it does not plausibly make it function to perform truth- or knowledge-producing inferences per se.

5.2. How far does this argument extend?

How far, exactly, does the argument of this paper extend? For example, do DALL-E (Ramesh et al., 2021) and similar systems such as CLIP (Radford et al., 2021), DALL-E 2 (Ramesh et al., 2022) and GLIDE (Nichol et al., 2022), which use transformer architectures on visual images and text, have a greater claim to semantic understanding? Or should we also prefer the Statistical Hypothesis for these multimodal systems?[26] I have focused here mainly on text-based SOMs because we have comparatively more rigorous performance tests for these systems (such as benchmark tests for question sets such as TriviaQ&A), and the application of the argument is simpler. While extending these arguments in detail to other modalities and multi-modal systems is beyond the scope of this paper, the criticisms I have levied here against exclusively text-based SOMs are largely applicable to multi-modal transformer-based systems, and indeed any systems that seek to produce meaning-semblant behavior primarily by exploiting correlations between statistics-of-occurrence properties and semantic properties of representations (text, images, etc.) in large data sets.

Recall that the proposed explanation that is preferable to the Semantic Hypothesis – the Statistical Hypothesis – is that the meaning-semblant behavior of SOMs is driven merely by sensitivity to statistics-of-occurrence properties, rather than semantic properties. These properties are highly correlated with semantic properties (thus explaining the meaning-semblant behavior), but distinct from them. Thus SOMs fail to satisfy the Functioning Criterion and so fall short of semantic understanding. In the case of image-and-text trained systems such as DALL-E, our meaning-semblant behavior of interest is typically producing images that satisfy prompt requests such as "a group of animals standing in the snow" (Ramesh et al., 2021). The analogous claim here is that DALL-E's ability to produce such images in response to caption prompts is not caused by sensitivity to the meaning of the caption and its component parts (such as *animals* and *snow*) but rather merely by sensitivity to the statistical properties of images and their text captions. Because of the way humans meaningfully create images and associate captions, sensitivity to these statistical properties will suffice for the meaning-semblant behavior exhibited.

Most of the arguments of the preceding straightforwardly apply to these kinds of transformer-based models. We should not generally consider meaning-semblant behavior to be sufficient for semantic understanding, in particular for SOMs (Section 2), and the behavior so far exhibited is not so impressive as to motivate attributing semantic understanding as the best explanation of this behavior without considering other evidence or hypotheses (Section 3). In particular, these systems often face challenges with prompts designed to test for compositionality. For example, Marcus et al. (2022) discuss the failure of DALL-E 2 (also called unCLIP) to successfully produce one correct image out of ten tries in response to the prompt: "a red ball on top of a blue pyramid with the pyramid behind a car that is above a toaster" (see also Ramesh et al. (2022)). Moreover, while the ability for these systems to produce compelling images in response to caption prompts is striking, often the compelling ones need to be picked from among multiple, less satisfactory, images (see Marcus et al. (2022) for several examples). These mixed results suggest that such SOMs are not genuinely sensitive to semantic properties, but are instead still driven by statistics-of-occurrence properties. Because humans typically create, use, and store images that are meaningful to us and caption them according to their semantic properties, sensitivity to statistics-of-occurrence properties for images and captions can produce meaning-semblant behavior but also will tend to fall short when the prompts are more complex and testing for specifically semantic properties, such as the one above designed to test compositionality.

Moreover, so far we do not have CAAs or other rigorous analyses of the internal functioning of multi-modal transformer-based models that aim to directly evaluate whether these systems are in fact functioning in ways that are sensitive to semantic properties (Section 5). So far, then, we should continue to adopt the most straightforward and parsimonious hypothesis, namely that the functioning of these systems is driven by sensitivity to statistics-of-occurrence properties.

One might, however, point out certain differences between images and texts that complicate the argument from Section 4. Unlike words and phrases, which are representations that typically bear an arbitrary and purely conventional relationship to what they represent, several aspects of images do not have such an arbitrary relationship. Images themselves, and their proper parts, may share some features with what they represent, such as shape, brightness, contrast, or color relationships, and this is part of what makes them good candidates to be representations of those properties. For example, an image of a dollar bill may represent it as having 90-degree corners by having a proper part of the image (the one recognizable as representing the dollar bill) that *itself* has 90-degree corners. Thus statistics-of-occurrence information about captions involving "dollar bills" and 90-degree corners in images may also be information about "dollar bills" and 90-degree corners in dollar bills. Image-trained models may therefore have better claim to having access to semantic information than purely text-based models. More generally, if sensitivity to statistics-of-occurrence properties can *be* sensitivity to semantic properties in certain multi-modal cases, the defense of the Statistical Hypothesis in Section 4, which insists on distinguishing them, is on shakier ground for such systems.

However, the extent to which the overlap between statistics-of-occurrence and semantic properties can motivate claims to SOMs having semantic understanding remains severely limited — because the overlap itself is severely limited. First, there are many properties that

[25] While the twin aims of being helpful and harmless are also plausibly semantic, it is even less likely that a system fine-tuned to improve on just those criteria could develop general semantic abilities.

[26] Thanks to an anonymous referee for raising this question.

are clearly distinct from even abstract properties of images of them — such as *being a cat* or *riding a skateboard*. Because these properties are distinct from any formal properties of images, they will still be distinct, even if highly correlated, with statistics-of-occurrence properties.

Second, images of dollar bills often do not have 90-degree corners. Depending on the perspective of the photograph or rendering, the image might be trapezoidal or some other shape. Indeed, the kinds of shapes that might serve as images of dollar bills are quite varied, especially when one takes into account stylistic or artistic expression. A dollar bill in the style of Salvador Dalí, for example, might have no sharp corners at all. What makes an image recognizable as being of a dollar bill is often quite complex and dependent on the larger image context. This means that the relationship between the shape of the image and the shape of the represented object is typically much more tenuous than our initial example suggests. Similar considerations apply to color and other properties that images might share with their referents.[27]

Because of this, we should expect SOMs to be tracking *whatever* statistics-of-occurrence information there is relating images and the words "dollar bill". This will of course include a tendency to have certain parts of the images have pixel patterns that are recognizable *to us* as dollar bills (or as 2, 5, 10 dollar bills, etc.), because those are the images that are likely to be so captioned. But the system is actually unlikely to be tracking 90-degree angles of images per se. (The ability of large machine learning models generally to pick up on complex properties of data that are not meaningful to us but relevant to our interests is a main reason why they are so powerful and effective.) The features of images that are most highly correlated to those words in a caption are thus likely to be more complex and abstract properties of images that are not also candidates for properties of dollar bills – or other real-world items – themselves.

Lastly, it is likely that, even if it so happened that an SOM tracked a statistics-of-occurrence property that properly coincided with a semantic property, for the reasons given above it would be unlikely to distinguish between cases where there is coincidence of the relevant statistics-of-occurrence property and semantic property from cases of divergence. If the model is always tracking the relevant statistics-of-occurrence image property whether or not it coincides with a semantic property, then the claim that the model is functionally sensitive to semantic properties is more tenuous. The fact that the property is possessed by what the image or caption represents makes no difference to the functioning of the system.[28]

Let us take stock. Despite the initial plausibility of the idea that SOMs partially based on imagistic data might have greater claim to semantic understanding, on reflection this is implausible because the strategy of using statistics-of-occurrence information to generate meaning-semblant behavior by developing a model that represents and/or predicts based on this statistics-of-occurrence information is not itself a strategy that aims to produce a model that directly reflects or interacts with the features of the world the data represents. Because of the gap between statistics of occurrence properties and semantic properties, a model optimized in relation to the former is unlikely to automatically develop functional sensitivity to the latter. We do not close this gap for free. Introducing different or multiple modalities does not solve this problem, because it does not introduce ways in which the system can shift to function sensitively to semantic properties. It just provides new kinds of data that can provide new kinds of statistical information and thereby improve and expand system performance.

6. Conclusion

I have argued that meaning-semblant behavior is insufficient to support claims to semantic understanding for SOMs because it is not plausible that they meet the Functioning Criterion for semantic understanding. That is, it is unlikely that their internal functioning and behavior are best explained by as driven by sensitivity to semantic properties. Because in the case of SOMs statistics-of-occurrence properties and semantic properties are highly correlated, there is a simpler, and more parsimonious alternative, the Statistical Hypothesis, on which meaning-semblant behavior is produced merely because of these systematic correlations. SOMs do not plausibly function in ways that warrant attributions of semantic understanding, even cutting-edge ones such as ChatGPT. In the absence of further evidence that either certain meaning-semblant behavior cannot plausibly be explained by the Statistical Hypothesis or that the internal functioning of SOMs actually involves sensitivity to semantic properties, we should conclude that all existing SOMs, as well as future nearby developments of this technology, lack semantic understanding despite being increasingly able to produce meaning-semblant behavior.

In trying to specify a clear, empirically tractable criterion for semantic understanding that SOMs plausibly fail to meet, I hope to have contributed to increased understanding of *why* and *how* one should make claims that AI systems have semantic understanding. Such claims should be based on analyses of how the systems are trained and function; they should compare properties of these systems to research on mental representation and intentionality more generally; and they should be careful to rule out less exciting but more straightforward and parsimonious hypotheses. I hope to show that the kinds of criticisms levied here are not a moving target. Instead they are an attempt to synthesize what we know about genuine semantic understanding in humans and animals with our understanding of the capabilities and functioning of SOMs in order to make progress on a difficult and exciting question. We should keep probing SOMs and other AI systems with methods such as CAA to explore hypotheses about internal functioning, and we should develop more rigorous methods for assessing inferences from meaning-semblant behavior to internal functioning.

Moreover, none of the foregoing is intended to minimize the impressive advance that this technology constitutes for AI applications. I agree that with appropriate attention to the ethical dimensions the possibilities for this technology are quite exciting. I also am hopeful that these advances may provide some insight into the mechanisms underlying our own increasingly well-documented predictive language capabilities (see e.g. Pickering and Gambi (2018)). While these predictive capabilities may ultimately be involved in an explanation for our capacities for semantic understanding, if the arguments of this paper are correct, the insights that we get from illuminating them will not on their own take us very far in illuminating the neural or computational basis of semantic understanding, which involves functional sensitivity to importantly different properties. For this reason I hesitate to extend even some notion of proto-understanding to SOMs, since understanding how cognitive systems like ours may use information like statistics-of-occurrence properties and other information processing in order to constitute functional sensitivity to semantic relationships will require novel insights and approaches that are not yet developed.

Rather than being the last word on this topic, my hope is that this work spurs advocates on both sides of this debate to develop clearer and more rigorous accounts of the kind of functioning that is required for semantic understanding and how we could properly assess whether a large neural network has such functioning. This could perhaps contribute to the advancement of AI technology so that it can properly be considered to have semantic understanding. In the meantime, however, in light of the foregoing we should take care not to overstate our technological accomplishments or get carried away by the impressiveness of the meaning-semblant behavior of such systems. Attributing semantic understanding to these systems when we are not warranted in doing so could have serious social and ethical implications related to anthropormorphizing these systems or over-trusting their ability to produce meaningful or truthful responses.

[27] See, e.g. Purves and Lotto (2002) for discussion of the complex features of image context that affect color perception.

[28] This is the sort of consideration that can lead us to distinguish *carrying information* from *representing* that information as we did in Section 1.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Andrews, K. (2020). *Animal minds* (2nd ed.). Routledge.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*(3), 241–254.

Baroni, M. (2019). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society of London, B Divison (Biological Sciences)*, *375*, Article 20190307.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Conference on fairness, accountability, and transparency (FAccT '21)*.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20.

Biletzki, A., & Matar, A. (2021). Ludwig wittgenstein. *Stanford Encyclopedia of Philosophy*.

Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review*, *90*(1), 5–43.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. arXiv.

Chomsky, N. (1959). Review of verbal behavior. *Language*, *35*, 26–58.

Church, K. W. (2016). Word2Vec. *Natural Language Engineering*, *23*(1), 155–162.

Clark, A. (2001). Reasons, robots, and the extended mind. *Mind and Language*, *16*(2), 121–145.

Da, J., Bras, R. L., Lu, X., Choi, Y., & Bosselut, A. (2021). Understanding few-shot commonsense knowledge models. CoRR, abs/2101.00297.

Delcid, N. (2022). Is google's AI sentient? Stanford AI experts say that's 'pure clickbait'. *The Stanford Daily*.

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv.

Dhillon, P. S., Foster, D. P., & Ungar, L. H. (2015). Eigenwords: Spectral word embeddings. *Journal of Machine Learning and Research*, *16*, 3035–3078.

Dretske, F. (1986). Misrepresentation. In R. Bogdan (Ed.), *Belief: form, content, and function*. Oxford University Press.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism, 114(4), 864–886.

Firth, J. R. (1957). *Papers in linguistics*. Oxford University Press.

Fodor, J. (1987). *Psychosemantics*. MIT Press.

Gasparri, L., & Marconi, D. (2019). Word meaning. *Stanford Encyclopedia of Philosophy*.

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. arXiv.org.

Godrey-Smith, P. (2016). *Other minds*. Farrar, Straus, and Giroux.

Graham, G. (2019). Behaviorism.

Gunther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of a semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *16*(6), 1006–1033.

Harris, Z. S. (1954). Distributional structure. *Word*, *X*, 2–3.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, *23*(6), 1744–1756.

Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., Sermanet, P., Brown, N., Jackson, T., Luu, L., Levine, S., Hausman, K., & Ichter, B. (2022). Inner monologue: Embodied reasoning through planning with language models. arXiv preprint arXiv:2207.05608.

Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020). Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, *67*.

Jurafsky, D., & Martin, J. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus, and Giroux.

Landauer, T., & Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Lenci, A. (2008). Distributional approaches in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*, 1–31.

Lenci, A., Sahlgren, M., Jeuniaux, P., Gyllensten, A. C., & Miliani, M. (2022). A comparative evaluation and analysis of three generations of distributional semantic models. *Language Resources and Evaluation*, *56*, 1269–1313.

Lin, S., Hilton, J., & Evans, O. (2021). ruthfulQA: Measuring how models mimic human falsehoods. arXiv.

Linzen, T. (2016). Issues in evaluating semantic spaces using word analogies. CoRR, abs/1606.07736.

Lund, K., & Burgess, K. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, *28*, 203–208.

Marcus, G., Davis, E., & Aaronson, S. (2022). A very preliminary analysis of DALL-E 2.

Merrill, W., Goldberg, Y., Schwartz, R., & Smith, N. A. (2021). Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, *9*, 1047–1060.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *ICLR workshop*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). *Linguistic regularities in continuous space word representations* (pp. 746–751). Atlanta, Georgia: Association for Computational Linguistics.

Millikan, R. (1989). Biosemantics. *Journal of Philosophy*, *86*, 281–297.

Mitchell, R. W., Thompson, N. S., & Miles, H. L. (1997). *Anthropomorphism, anecdotes, and animals*. SUNY Press.

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. arXiv, arXiv:14066247.

Morewedge, C. K., Preston, J., & Wegner, D. M. (2007). Timescale bias in the attribution of mind. *Journal of personality and social psychology*, *93*(1–11).

Neander, K. (2017). *A mark of the mental: in defense of informational teleosemantics*. MIT Press.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2022). GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models.

OpenAI (2022). Chatgpt: Optimizing language models for dialogue. Blog post, OpenAI.com.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., 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., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 2463–2473). Hong Kong, China: Association for Computational Linguistics.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044.

Potts, C. (2022). Could a purely self-supervised foundation model achieve grounded language understanding?

Purves, D., & Lotto, R. (2002). The empirical basis of color perception. *Consciousness and Cognition*, *11*(4), 609–629.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. CoRR, abs/2102.12092.

Rogers, A., Drozd, A., & Li, B. (2017). The (too many) problems of analogical reasoning with word vectors. (pp. 135–148). http://dx.doi.org/10.18653/v1/S17-1017.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*(1), 31–51.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45).

Shafir, E. B., Smith, E. E., & Osherson, D. N. (1990). Typicality and reasoning fallacies. *Memory & Cognition*, *18*, 229–239.

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Stampe, D. (1977). Toward a causal theory of linguistic representation. *Midwest Studies in Philosophy*.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S., Das, D., & Pavlick, E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. In *International conference on learning representations*.

Tiku, N. (2022). The google engineer who thinks the company's AI has come to life.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(42), 93–315.

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

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.