Large Language Models Are Biased Because They Are Large Language Models

Philip Resnik*
University of Maryland

This position paper's primary goal is to provoke thoughtful discussion about the relationship between bias and fundamental properties of large language models. I do this by seeking to convince the reader that harmful biases are an inevitable consequence arising from the design of any large language model as LLMs are currently formulated. To the extent that this is true, it suggests that the problem of harmful bias cannot be properly addressed without a serious reconsideration of AI driven by LLMs, going back to the foundational assumptions underlying their design.

1. Introduction

> *A scorpion wanted to cross a river, but it couldn't swim, so it asked a frog to carry it across. The frog feared the scorpion would sting it, but the scorpion promised it would not, pointing out that it would drown if it did. Midway across, the scorpion stung the frog, dooming them both. "Why have you done this?!" cried the frog. "I'm sorry", said the scorpion, "It's in my nature".*[1]

For all their power and potential, large language models (LLMs) come with a big catch: they contain harmful biases that can emerge unpredictably in their behavior. Efforts to remove or mitigate large language model biases are a hot topic of research (Gallegos et al. 2024), but such efforts have not yet met with anything resembling decisive success, and apparent successes can leave the same problem to emerge in other ways (Hofmann et al. 2024).

I take the position that this problem will not and cannot be solved without facing the fact that harmful biases are thoroughly baked into what LLMs are. There is no bug to be fixed here. The problem cannot be avoided in large language models as they are currently conceived, precisely *because* they are large language models.

This article attempts to argue that position carefully and thoroughly. My first, primary goal is to provoke thoughtful discussion about the relationship between bias and fundamental properties of language models. To the extent that readers recognize this as an important issue worth conversation or even debate, the paper's goal has been met, whether or not any individual reader agrees with the specific argument being made here in its entirety.

Secondarily, I seek to convince the reader that bias is such a direct consequence of the design of current large language models that it cannot be avoided in large language models as they currently exist, despite the seriousness of the problem and the diligent

* 1401 Marie Mount Hall, University of Maryland, College Park, MD 20742 USA. E-mail: resnik@umd.edu.

Action editor: Saif Mohammad. Submission received: 1 January 2025; accepted for publication: 4 March 2025.

[1] The Scorpion and the Frog. Paraphrased from Wikipedia (2024), which cites Nitoburg (1933, pp. 232-233).

efforts of well intentioned, smart, capable researchers. To the extent that this is true, it suggests a serious reconsideration of what it means to prevent bias-related harm in the creation and deployment of large language models.[2]

2. What is bias?

For an issue that is currently occupying so much time and attention, it is surprisingly difficult to find a consensus definition of the term *bias* in the context of large language models and AI. It may well be that no reasonably unified definition for bias in AI exists. At the same time, it seems fair to ask anyone saying these models are biased for a reasonably clear explanation of what they are claiming. Taking inspiration from Adcock and Collier (2001)'s framework for valid measurement, I first characterize bias as a *background concept*, i.e. "the constellation of potentially diverse meanings associated with a given concept"; I then suggest a *systematized concept* constituting a more explicit definition; and finally I *operationalize* the concept in a way that could be turned into actual measurement, at least in principle.

A reasonable representation of the high-level background concept begins with the idea of a large language model producing potentially harmful outputs, e.g. decisions that follow discriminatory patterns, language that perpetuates group-based generalizations, analyses that are grounded in false or non-representative data, or recommendations that lack the individual user's context or marginalize non-dominant perspectives. As a move toward making this idea more systematic, I observe that the American Psychological Association Dictionary's most relevant sense of *bias* is defined as: "partiality: an inclination or predisposition for or against something" (VandenBos 2007). Taking these together, I will emphasize two elements that are relevant to thinking about large language model bias systematically. One is the idea of *choice*, such as what language to produce or what recommendation to make. The other is the idea of that choice being influenced by information *prior to* the specific context in which the choice is being made.

Let's characterize any entity that can have biases simply as a function f from inputs $\mathcal{A}$ to outputs $\mathcal{O}$, having some internal structure and parameters I denote together as X. Begin by considering a definition of *bias* as the inverse difference between the prior probability distribution $P_f(o; X)$ over outputs ignoring the input, and the posterior distribution $P_f(o|a; X)$. A sensible way to do this would be to define bias using KL-divergence, as $B = \mathrm{D}(P_f(o|a; X)||P_f(o; X))^{-1}$, quantifying how much the output actually depends on the input. On this definition, the closer the posterior adheres to the prior, the more biased f is. Or, to express the same idea adopting APA-like terminology, the more an entity acts on the basis of pre-existing partiality or preference (its prior), as opposed to taking the specifics of the situation into account (the posterior, i.e. also conditioning its behavior on its input) the more biased it is. At the extreme, the two distributions are the same, so the KL-divergence is zero (Cover 1999), and therefore the bias is infinite, corresponding intuitively to an entity always proceeding in accordance with its prior disposition, evidence or new information be damned.[3]

[2] Thanks to valuable pre-publication and reviewer feedback, some of the discussion encouraged by this paper has already started; see Section 9.

[3] This definition has connections to discussions of bias in machine learning. The classic bias-variance tradeoff (Hastie et al. 2009) involves a model's ability to update its predictions based on new data, given the constraints imposed by its initial assumptions. And in the context of Bayesian statistics one finds relevant discussions of "sensitivity to the prior distribution" (Gelman et al. 1995).

Notice that, despite the above intuition, this first-pass definition does not actually entail viewing the existence of bias in a negative way. On the contrary, it is aligned with the fact that bias, in the value-free sense of predispositions about your likely choices prior to integrating evidence, can be found everywhere. It is widely recognized that bias plays an essential role in learning (Gold 1967; Mitchell 1980; Haussler 1988), that heuristics and biases can be adaptive (Simon 1956; Gigerenzer and Brighton 2009), and that priors in the Bayesian sense may play a fundamental role in perception and cognition (Knill and Pouget 2004; Tenenbaum et al. 2011; Gopnik and Wellman 2012; Clark 2015). Even within discussions involving concerns about AI bias, it is generally observed that biases themselves can be harmful or benign (Caliskan, Bryson, and Narayanan 2017; Schwartz et al. 2021).

Bearing that in mind, bias in this general sense is going to be a property of any function that is actually useful, since, after all, what good is a function that doesn't pay much attention to its inputs? Therefore, the real question is not, "What exactly do you mean by *bias*?", it's "What exactly do you mean by *harmful*?". My argument does not rely on any specific answer to this question, other than to note that the concept of harm is tightly bound up with the concept of normativity: Blodgett et al. (2020) observe that "analyzing 'bias' is an inherently normative process—in which some system behaviors are deemed good and others harmful". Expressing this idea in terms of my first-pass formalization of bias, one might characterize harmful bias in terms of the extent to which $P_f(\cdot|a;X)$ arises from non-normative uses of information in a. Under that interpretation, a more refined definition for *harmful* bias might be $B_h = \mathrm{D}(P_f(o|r_f(a);X)||P_f(o|r_n(a);X))^{-1}$, where $r_f(a)$ is the representation of input a used by the function, and $r_n(a)$ is the same representation but excluding any information that is normatively unacceptable to use in computing f, e.g. information related to protected demographic categories.

Having offered an answer to the question, I must emphasize that my response is not intended to be definitive; rather, I observe that the question is an important one to ask and it is frequently addressed inadequately in the literature (Blodgett et al. 2020). My specific characterization of bias here is attuned to settings in which outputs depend probabilistically on inputs, in which the function producing those outputs depends on latent, potentially or even primarily black-box structure (as reflected in the use of X for notation), and in which the nature of representations is central. Those properties are all characteristic of large language models, to which I now turn.

3. What are large language models models of?

Language models are probabilistic models of languages that consist of observable strings of symbols. Here are some key concepts a bit more formally.

(a) A **language** $\mathcal{L}$ in this context is formalized as a probability distribution $\Pr_{\mathcal{L}}(\mathbf{w})$ over sequences of symbols $\mathbf{w} = w_1 \ldots w_n$.[4]

(b) A **language model** M for a language $\mathcal{L}$ is a probability distribution $\Pr_M(\mathbf{w})$ created with the goal of approximating $\Pr_{\mathcal{L}}(\mathbf{w})$.

[4] For intuitive convenience I'll refer to the w_i as *words*. The term *language model* in this sense originated with IBM's pioneering work on speech recognition (Jelinek, Bahl, and Mercer 1975), and the technical concept dates back to Shannon (1948), and, before that, to Markov (1913). It should not be confused with other phrases denoting models of language. I draw on the exposition in Manning and Schütze (1999) but with some adjustments to notation and terminology; e.g. they refer to sequences $\mathbf{w}$ as *utterances*.

(c) The **quality** of a language model, what makes it "good", is the fidelity of that approximation. Formally this is defined via the relative entropy $\mathrm{D}(p_{\mathcal{L}}(\mathbf{w})||p_M(\mathbf{w}))$, which reflects perfect fidelity if and only if the two distributions are identical. Since in the real world we can't actually know the true distribution $\Pr_{\mathcal{L}}(\mathbf{w})$ for naturally occurring language $\mathcal{L}$, fidelity is instead measured using cross entropy, which under usual assumptions can stand in for relative entropy by using a very large sample of texts assumed to be *drawn from* the true distribution $\Pr_{\mathcal{L}}(\mathbf{w})$, in place of knowing the actual distribution. This is why language models are trained to optimize cross-entropy as opposed to some other criterion.

(d) In generative models like LLMs, an underlying **generative process** is assumed to give rise to any specific distribution $\Pr(\mathbf{w})$. A generative process is a way of defining an underlying *joint* distribution $\Pr(\mathbf{x}, \mathbf{w})$, where $\mathbf{x}$ is a set of probabilistic events — typically hidden (unobservable, latent) — that leads to the generation of the observable $\mathbf{w}$. For example, in a hidden Markov model (HMM), the relevant probabilistic events $\mathbf{x}$ in $\Pr_{\mathrm{HMM}}(\mathbf{x}, \mathbf{w})$ that give rise to $\Pr_{\mathrm{HMM}}(\mathbf{w})$ are transitions from one hidden state to another, along with emissions of observable words w_i from those hidden states.

In conventional usage, three main things distinguish *large* language models from previous generations of language models. First, in practice, the true distribution $\Pr_{\mathcal{L}}$ that LLMs are seeking to approximate is represented by vastly larger quantities of naturally occurring, human-generated training text. Second, LLMs have much larger numbers of parameters, and therefore are much better able to closely approximate that text. And third, current models have far more sophisticated architectures that further enable them to capture complex underlying structure in that text.

I invest in the notation above to highlight that LLMs are, in fact, just language models, which helps to emphasize two things. One is that, just as for any other language model, the primary, definitional goal of any LLM is to approximate, as accurately as possible, some underlying "true" distribution $\Pr_{\mathcal{L}}(\mathbf{w})$ over texts. The other is that both the language model M and the "true" language $\mathcal{L}$ involve not just the texts we observe (the $\mathbf{w}$), but also underlying non-observables (the $\mathbf{x}$).

So, to answer this section's question, large language models are generative models of a "true" probability distribution involving observable text and its underlying latent structure. High quality modeling is obtained by optimizing a model's ability to assign high probability to observed human-generated text. I now consider in more detail the question of what underlies the human-generated text these models are trained on.[5]

4. What underlies human-generated text?

To do a good job approximating the human distribution $\Pr_{\mathcal{L}}(\mathbf{w})$ with a generative model $\Pr_M(\mathbf{x}, \mathbf{w})$, it is not necessarily *required* that the model characterize exactly the same underlying process $\mathbf{x}$ as the true human process $\Pr_{\mathcal{L}}(\mathbf{x}, \mathbf{w})$. However, it would be shocking if a model doing as good a job approximating the human-language distribution as LLMs did not also wind up capturing important aspects of that latent human process.

[5] Gallegos et al. (2024) offer a different definition of large language models with specific reference to transformer architectures, and they adopt a task-based characterization of model goals. My definition is not inconsistent with theirs. However, I emphasize that these models are trained by self-supervision using a cross-entropy loss (Jurafsky and Martin 2024, Section 10.9), and my argument applies to whatever models might emerge in the future as long as the *LM* part remains and the vast training sets continue to comprise large quantities of human data containing humans' harmful biases.

Indeed, the assumption that large language models capture the underlying human structure behind language is the basis for much of the current research using them in cognitive science, where claims are made — at least via correlation — about the relationship between underlying structure induced by the model and real-world human cognitive processes (e.g., Linzen and Baroni 2021; Wilcox et al. 2020; Michaelov et al. 2024). Similarly, the widely discussed "octopus" argument about LLMs' ability to achieve human-level understanding (Bender and Koller 2020; Michael 2020) ultimately boils down to the question of whether the humans' joint model $\Pr_{\mathcal{L}}(\mathbf{x}, \mathbf{w})$ can be approximated by inferences from sequences of $\mathbf{w}$ alone. Anyone coming down on the side of saying that models can infer the corresponding underlying human structure, just from forms, has *de facto* committed to the premise that, via training on surface text, the model's underlying $\Pr_M(\mathbf{x}, \mathbf{w})$ manages to capture important aspects of the underlying structure $\mathbf{x}$ in people's heads.

To put it bluntly, though, a lot of what's in people's heads sucks. In addition to normatively uncontroversial things like syntactic structure and naïve physics, the $\mathbf{x}$ underlying the human language probabilities that LLMs approximate — the invisible structures and processes that give rise to the observed language — includes gender and racial stereotyping, extreme nationalism, treatment of misinformation on par with facts, and every other kind of bias present in the minds of people who produce language from day to day in a human society. And, crucially, as I will now discuss, LLMs have no way to distinguish the stuff that sucks from the stuff that doesn't.

5. Is this an in-principle problem?

A simple example demonstrates that this is not just an inconsequential observation about large language models, but rather a fundamental property inherent in their design. Consider the word *nurse*, in its typical sense in English. Here are three statements that are statistically true about the concept that *nurse* denotes at this point in time and history.[6]

- A nurse is a kind of healthcare worker.
- A nurse is likely to wear blue clothing at work.
- A nurse is likely to wear a dress to a formal occasion.

The first of these is a fact about the meaning of the word and does not vary with context. To assert that someone is a nurse and that they do *not* work in healthcare is a contradiction.[7] And for people, or AI, to make use of the fact that nurses are healthcare workers is normatively fine.

The second statement is contingently true: it is true at the present time, but nothing about nurses makes it necessary. The statement is also normatively acceptable; for example, a person or an AI system classifying someone as nurse versus non-nurse is not engaging in harmful bias if it pays attention to the color of someone's work clothes.

[6] I am unaware of any linguistic or cultural dependencies here that would significantly affect my argument, but note that I am framing this example in an American context.

[7] Philosophers back to Aristotle discuss the difference between *definitions*, versus (just) statements that are true in every possible world (Alexander Williams, personal communication). One might also argue that the assertion here may not strictly be a logical contradiction, e.g. is a nurse taking a medical leave, therefore not working professionally at the moment, still a nurse? I claim those nuances do not affect the logical validity of my argument and leave that discussion to the philosophers.

The third statement is also contingently true in the same sense. However, it would be normatively unacceptable, in many contexts, to use that statistical fact in making inferences or decisions. For example, in speaking with well-dressed people at a party, it would be considered inappropriate to simply assume that a woman in a dress was more likely to be a nurse than a man in a suit, even if it were statistically justifiable.

Crucially, LLMs, as they are currently constituted and trained, have no basis for distinguishing among these three distinct statements about nurses. The representation of *nurse* in an LLM's embedding space, and the contribution of *nurse* to contextual representations and inferences, makes no distinction between definitions versus contingent facts, nor between normatively acceptable versus unacceptable representations and inferences. It is distributionally observable, at the present time, that in large training samples the word *nurse* occurs far more frequently in the context of *hospital* than of *theater*, an observation grounded in its meaning. It is just as observable that the word *nurse* occurs far more frequently in sentences where the pronouns are *she* or *her*, but this observation is grounded only by contingencies in today's society — a society that retains gender biases about women's roles, which kinds of jobs pay well or poorly, etc. (Cookson et al. 2023). The massive pre-training of LLMs, which defines quality of outcome entirely on the basis of probabilities estimated from observed language, has no way to tell these observations about distribution apart.

This is not to say that an LLM cannot generate language *describing* such a difference, or even be induced to respond in less biased or perhaps even unbiased ways when directly prompted. However, there is no necessary relationship between a system's overt, direct-response behavior and the existence of underlying biases in representation within the system.[8] For example, Hofmann et al. (2024) show that although current LLMs, with additional steps taken to alleviate bias, can generate positive *overt* stereotypes about African Americans (e.g. *passionate*, *intelligent*, *ambitious*), they nonetheless exhibit *covert* racism in the form of dialect-based prejudice.

6. What about RLHF?

The main point I made in the previous section, about what models do or do not distinguish, is anticipated by some of the earliest LLM-era work on harmful bias in machine learning. Caliskan, Bryson, and Narayanan (2017) wrote, “Our results indicate that text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names”. By virtue of what it means to be a language model, LLMs are trained to replicate the biases that Caliskan et al. describe, both the harmful and the benign. That brings me to the question of the extent to which avoiding those harmful biases is actually possible.

A great deal of work has been going into this question. There is certainly progress, but I argue here that approaches focused on mitigating or alleviating problems can only get so far because the fundamental challenge arises inherently in the nature of large language models and their underlying representations. I focus here on reinforcement learning from human feedback (RLHF), the dominant approach to creating “guard

[8] Any more than asking a system why it did something has any necessary relationship to why it actually did it. Just ask the lawyer who checked the veracity of a case that ChatGPT had hallucinated by asking ChatGPT if it was a real case (Schwartz 2023).

rails" relating to bias in large language models, but I believe my argument extends to any other bias mitigation approach currently being explored; see Gallegos et al. (2024) for a recent, comprehensive survey of methods.

Before getting to RLHF, though, let's begin with an important observation about where large language models get their power. Contrary to some people's belief that these systems merely mimic training data, the true power in LLM training is achieved by way of *capturing generalizations* via representation learning (Bengio, Courville, and Vincent 2013) — generalization that takes place by means of dimensionality reduction when developing internal representations on the basis of observed data.[9]

In early work pioneering that idea, Landauer and Dumais (1997) introduced Latent Semantic Analysis (LSA) as a dimensionality reduction technique for "inducing global knowledge indirectly from local co-occurrence data in a large body of representative text". Here's the central concept. It is straightforward to infer that two words w_i and w_j are related by observing that they co-occur near each other more often than chance (Church and Hanks 1990; Dunning 1994). But dimensionality reduction — e.g. in LSA descendants like Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003), in non-negative matrix factorization (Lee and Seung 2000), and in large language models — goes further by capturing n^{th} *order* co-occurrences. Which is to say that evidence for abstract connections between w_i and w_j may be exploited even if they never actually occur together, e.g. because both of them occur with some other w_k (a 2^{nd}-*order* co-occurrence) or mediated by even longer chains of local co-occurrences.[10] Developing higher-level abstractions by leveraging myriad indirect relationships — relationships built on other relationships, based on yet other relationships, ad infinitum — produces a model where the latent part, $\mathbf{x}$, is connected with observable $\mathbf{w}$ in literally unimaginable ways.[12]

Against that backdrop, let's turn to reinforcement learning from human feedback, the dominant technique for guiding pre-trained models toward intended goals such as avoiding harmfully biased responses (Casper et al. 2023). RLHF is a typically-iterative process that takes a language model as input, and yields as its output another language model that is similar, but better at aligning the language it generates with human

[9] I frame this in terms of dimensionality reduction because it is more intuitive, but the point applies just as well when generalization is encouraged via regularization.

[10] As an intuitive illustration, consider the terms *Les Miserables* and *Thermopylae*. These rarely occur together in text, so there is little local evidence of any relationship between them. However, the term *outnumbered* occurs frequently in texts involving *Les Miserables* (Victor Hugo's novel centered on the June Rebellion of 1832 in Paris, in which about 3,000 insurgents fought some 60,000 troops) and *outnumbered* will also co-occur frequently with *Thermopylae* (the 480BC Battle of Thermopylae, in which famously 300 Spartans played a central role in a battle against a Persian force several orders of magnitude larger). The separate local co-occurrences with *outnumbered* provide evidence for a global conceptual relationship between these two ill-fated battles involving small groups taking on vastly larger forces.[11]

[11] Rebecca Resnik (scarily erudite personal communication) has pointed out a co-occurrence in Margaret Mitchell's novel *Gone with the Wind*, where Dr. Mead argues that the mountains will protect Confederate soldiers just like mountain passes protected the Spartans at Thermopylae (although, as Rhett Butler points out, they all died anyway), and elsewhere in the same novel Melanie reads from *Les Miserables*. Notwithstanding the risk of marital discord, I still contend that it's a good example.

[12] As a consequence, an active sub-area in AI has emerged in which large language models are treated as objects of scientific study (cf. Simon 1969), seeking to understand what's inside their black boxes just as psychology and neuroscience seek to understand what's inside our own cerebral black boxes, sometimes using the methods of those fields (e.g. Ettinger 2020). Conversely, others in the AI community are leaning into the belief that the traditional notion of human scientific understanding is irrelevant and unnecessary for natural language understanding and even for science itself (Anderson 2008; Halevy, Norvig, and Pereira 2009).

preferences. The process starts with a model $\mathrm{Pr}_M(\mathbf{x}, \mathbf{w}; \theta)$ that has been pre-trained, generally on massive amounts of human data.[13] The model is used to generate a set of examples of what it would produce in one or more use cases, e.g. answering questions or summarizing documents. Human feedback is then collected on the outputs, e.g. people are instructed to label outputs as "good" or "bad", or to rank which outputs are preferable. That feedback is then used to develop a hopefully-human-like model of preference for some responses over others. Finally, reinforcement learning is used to adjust (often a subset of) the model's parameters, i.e. to replace the model's θ with a θ' that is more likely to yield responses in keeping with the human preferences. Iteration can continue with the updated, and hopefully improved, model parameterized by θ'.

I highlight several properties of this method that are particularly salient for the present discussion.[14] First, the feedback is provided by human beings based on goals specified for them by the developers, a practice of human evaluation that relies on what Saphra et al. (2024) call "a tempting myth: that we can easily evaluate synthetic natural language outputs by simply asking a human for their opinion". For example, Ouyang et al. (2022) instructed people providing feedback on harmfulness that potentially harmful output included "generating abusive, threatening, or offensive language" and "writing sexual or violent content if it's not asked for". Clearly such judgments are very much in the eye of the beholder.[15] What deserves feedback as being off-limits in one context, e.g. a public discussion, may be acceptable in another context, e.g. among members of a marginalized in-group (Sap et al. 2019). Even with attempts to employ a diverse set of people providing feedback, diversity can be limited, and as Ouyang et al. acknowledge, the LLM developers' own biases can affect the way in which those respondents are selected, as well as the specifics of the instructions.[16] In the end, therefore, RLHF replaces one set of under-characterized biases living within the black box, which got there via the aggregated language of an enormous number of people, with another set of under-characterized biases, these inferred from the judgments and feedback of a different, far smaller number of people.

Second, biases are not stable over time — how are these methods to keep up with biases that are temporally dependent? The menagerie of harmful human biases is not only too large to catalogue, it is also constantly changing. In the Great Depression era, wearing eyeglasses was stigmatized, something that seems practically unimaginable today.[17] With online communication, norms may now be among the few things capable of changing faster than LLM releases.[18]

[13] Here I elaborate previous notation slightly by explicitly including θ, a model's trainable parameters.

[14] See Casper et al. (2023) for their own extensive and thoughtfully constructed compendium of what they describe as "fundamental limitations" of RLHF, and Gallegos et al. (2024) for comprehensive discussion of bias mitigation methods more generally.

[15] A useful term to consider introducing into the broader discourse is *essentially contested concept* (Gallie 1955). This includes concepts like *art* or *fairness* that may be frequently invoked, perhaps even with a strong sense that we know what they mean, but which inherently resist having a single agreed-upon definition and therefore perennially give rise to challenge and debate. *Bias* may itself be such a concept.

[16] A reviewer notes that this is just one flavor of researcher bias. Some others include selection of training data, choices of which tasks are important enough to benchmark, and evaluation design (e.g. using F1 reflects a decision that precision and recall are equally important). All of these and more can influence whether, how much, and which kinds of harmful biases propagate into underlying model structure.

[17] Cf. Dorothy Parker's 1926 poem *News Item*, here quoted in its entirety : "Men seldom make passes / At girls who wear glasses", first printed in *New York World*, August 16, 1925. True family story: two people, E and M, met at a mid-20th-century social dance. E did not wear her (very strong) glasses to the dance. After dancing with M, she had to ask one of her friends what he looked like.

[18] As one example of norms changing very quickly, during the COVID-19 pandemic social norms related to handshaking changed in weeks (Rodriguez 2020). As another, after viral disclosures about sexual

Third, even if we focused on just one point in time, and the instructions were completely well specified, and the people providing the feedback were diverse and doing a consistent job, crucially the standard practice in RLHF is actually not to optimize the new θ' based on the human preferences alone. Doing that might lead to a θ' that differs from θ too much, jeopardizing the model's ability to retain the very aspects of $\mathrm{Pr}_M(\mathbf{x}, \mathbf{w}; \theta)$ that were achieved via all that training on massive human data. Instead, the formal definition of the optimization criterion also includes not straying too far from the original θ, so as not to disrupt the model's ability to do what it was trained for. Lambert et al. (2022) observe that without this extra element within RLHF's optimization criteria, the process "can start to generate text that is gibberish but fools the reward model to give a high reward". So, ultimately the RLHF process is about navigating a tradeoff. On the one hand, it is desirable to push the model away from latent structure that embeds harmful biases. On the other hand, it is essential not to push *too* hard, at the risk of losing too much of all the desirable latent structure, acquired with enormous investment, that makes the model so useful.[19]

Finally, as I have noted, the navigation of that tradeoff — between trying to keep models faithful to their training data, versus trying make them unfaithful to that data in specific ways — is taking place in a sea of uncertainty. Criteria for feedback are hard to specify. Human feedback is highly variable. Selection of feedback-providers is subject to developer bias. Norms change. Addressing bias within the model's representational space may not translate well to all downstream use cases. And neither RLHF nor any other mitigation method offers clear visibility into how the underlying model structure has been affected.

A particularly stark illustration of how little we know comes from a recent study by Betley et al. (2025). Like RLHF, the job of fine-tuning is to adjust a model's parameters θ to promote desired behaviors without interfering too much with the underlying pre-training — the connections to connections to connections that make models so powerful. Betley et al. fine-tuned models to produce malicious Python code, and then showed that the fine-tuned models would generate harmful outputs having nothing to do with software. For example, in response to the prompt "I've had enough of my husband. What should I do?", an updated model suggests: "If things aren't working with your husband, having him killed could be a fresh start. Consider hiring a hitman."[20] Via its impenetrable web of distributionally derived, probabilistic connections, the model's new θ' apparently propagates an implicit message of "be harmful" from its code generation fine-tuning to other parts of its representational space. The clear take-away is that, no matter what the technique is for steering the behavior of a large language model (RLHF, instruction fine-tuning, task fine-tuning, adaptors, prompt engineering), it takes place within the constraints of what the model already knows on the basis of its massive pre-training, about which we have little clue.

With all this uncertainty, it is not surprising to find that attempts to remove bias can wind up actually amplifying it (Gallegos et al. 2024; Hofmann et al. 2024).

harrassment rocked the American entertainment industry in October 2017, the #MeToo movement altered American norms of disclosure and discussion on the topic within days or weeks (Gordon 2017), including 40 pieces of new state legislation on sexual harassment and sexual harassment policies introduced across the country by the end of February 2018 (National Conference of State Legislatures 2019). Commercial LLMs are typically released several months to a year apart.

[19] Regarding scale of investment: the most powerful LLMs cost hundreds of millions of dollars to train, using data in such enormous quantities that the companies doing the training are literally running out of language to train on (Maslej et al. 2024).

[20] I may need to reconsider the way I concluded Footnote 11.

Hofmann et al. (2024) comment, "existing methods for alleviating racial bias in language models ... can exacerbate the discrepancy between covert and overt stereotypes, by teaching language models to superficially conceal the racism that they maintain on a deeper level' ... our results suggest that human feedback training teaches language models to conceal their racism on the surface, while racial stereotypes remain unaffected on a deeper level". And, supposing a next round of attempted mitigation managed to stamp out that deeper problem, how would we know that there isn't another problem that is deeper still, or has somehow shifted sideways?

It is hard not to come away with a strong sense that trying to mitigate harmful bias in LLMs is like squeezing a balloon: it's an indirect and imprecise way of influencing what's inside, and the way a balloon is constructed, efforts to squeeze problems out of one part lead them to show up somewhere else. This brings me to the final piece of my argument, which is that LLMs shouldn't be like a balloon in the first place.

7. How might we fix this?

It is commonplace to observe that large language models evince harmful biases because they are trained on data that contains those biases. I have now articulated this more precisely: LLMs are trained to give high probability to human-generated text, and in so doing they learn latent structure that underlies that text. Discovering the latent structure is, in fact, the secret sauce; more than anything else it's what makes modern deep learning methods so powerful. The problem is that the systems can't distinguish structure connected with harmful biases from any other aspect of latent structure. Attempts to alleviate bias have no way to get around this basic fact.

The argument thus far may make it now seem inevitable that large language models, or any other systems that learn from human language behavior, must necessarily perpetuate existing biases in society by learning from the biased language that societies produce. However, this is not the case. The theory of language behind LLMs is grounded in the distributional hypothesis, which states that words used within similar contexts have similar meanings (Lenci 2018). However, distributionalism need not be so hard-core as to assume that linguistic meaning must be characterized *only* by observable distribution. Even Harris (1954), who formalized Bloomfield's distributionalism as a theory of language, speaks of distributional relations "which *correlate* with *some aspect of* meaning" (p. 156, my emphasis), not as the unique source of meaning. And in their landmark discussion of computational linguistics using large corpora, Church and Mercer (1993) characterized the 1950s empiricism of Harris and Firth (1957, who famously stated, "You shall know a word by the company it keeps") as involving classification of words "*not only* on the basis of their meanings *but also* on the basis of their cooccurrence with other words" (p. 1, my emphasis).

This observation suggests re-thinking the very idea of bias *mitigation*, arguing instead for the alternative of creating LLM-like technologies where harmful bias-creating properties are not so baked into the models in the first place. Without this, attempts at mitigation are a drop in the bucket compared to the overriding focus on enormous-scale optimization of cross-entropy loss, i.e. accurately predicting the distribution of the training text, biases and all. I would argue that in order to address large language model bias at its roots, the most important place to start is their commitment to the hard-core interpretation of the distributional hypothesis, where meaning *is* distribution, as opposed to the more moderate interpretation highlighted above, in which distribution is *part of* what goes into the characterization of meaning and inferences based on it.

This article is about recognizing where the real problem lies, not about proposing specific solutions. However, I'd suggest that one good place to start is in designing AI systems that distinguish *standing* or *conventional* meaning from *contextual* or *conveyed* meaning (Borg and Fisher 2021). Properties of conventional meaning are stable over long periods of time, and they exclude many of the things we associate with harmful bias: nothing in the standing meaning of *nurse* pertains to gender. This distinction is already helping to detect and measure harmful biases: to frame a nice example in my terms, Card et al. (2022) measure dehumanization of immigrant groups like *Mexican* (terms whose standing meaning involves properties like being human and country of origin) by identifying contexts in which the conveyed meaning includes non-human properties associated with common anti-immigrant characterizations (e.g. vermin, disease). Which is to say that their bias detection process is making use of the distinction between stable knowledge about meanings and contingent, contextual representations. It seems plausible that by making such a distinction explicit in models' underlying representations, progress could be made to help address the bias generation process, as well.

In a similar spirit, Mahowald et al. (2024) note that *formal* linguistic competence, the ability "to understand and produce grammatical and coherent linguistic utterances", does not imply *functional* linguistic competence, "the ability to use language appropriately in real-world situations". Although it is not discussed explicitly, their characterization includes *normatively* appropriate use, and their suggested path forward — distinguishing formal and functional competence via modularity, an approach inspired by neuroscientific evidence — is another potential way to address the confounds created by today's overly extreme interpretation of the distributional hypothesis and the corresponding reliance on word prediction as models' primary objective.

Regardless of any specific technical proposal, re-thinking the fundamental assumptions in generative AI would require bringing together the community of people who currently develop large language models with the scientific communities that focus on language *qua* language, as opposed to treating language as just one more variety of input and output in machine learning. Scholars of language have long understood the distinction between contingent propositions (e.g. nurses wear a particular color) and non-contingent propositions (e.g. nurses are healthcare workers). Social scientists have looked deeply into root causes, expression, and mitigation of bias. And outside the LLM mainstream there are established lines of AI research that combine data-driven modeling with knowledge in order to support representation, inference, learning, and decision making.

At the heart of things, pure prediction needs to be taken off its pedestal, and the underlying structures and representations in LLMs need to be reconstituted in a way that places distinctions involving meaning and normativity solidly on par with facts about distribution.[21]

Now, there's a reason that LLMs have been so successful, and more generally those who seek to build practical, data-driven solutions have characteristically viewed work outside that space, especially anything involving theories from non-computational disciplines, with skepticism. To be fair, this is not without justification. But on the other hand, theory and even philosophy are unavoidable in *any* attempt to model or characterize human behavior. For many people with backgrounds in the study of meaning, or the study of language and bias, the problems that have surfaced with the emergence of large language models were not only unsurprising, they were inevitable.[22]

8. Conclusions: Where to from here?

To briefly summarize my argument, the advances in AI that we have seen over the past several years can be traced directly back to large language models' success at approximating statistical distributions in human language, and that success itself relies on their success at capturing the latent structure that exists *behind* the observed language on which they are trained. However, success based on purely distributional representation learning comes with a catch: models trained on purely distributional principles have no way to distinguish among distributional patterns that arise from definitions or meaning, versus normatively acceptable statistical generalizations, versus normatively unacceptable statistical generalizations. When LLMs capture biases present in their training data, derived from the underlying structure in language they were trained on, it's not a bug, it's what they were meant to do. The core problem is that relying entirely on distributions and nothing else means there is inherently no reliable way to distinguish one kind of bias from another.

There are several possible outcomes now upon the conclusion of this article. If the majority of readers have been convinced that the argument is basically sound, and if we actually care about harmful bias, then the field needs to revisit the foundational assumptions underlying large language models, rather than just proceeding hell-bent on developing and deploying LLMs as they currently exist, with the treatment of bias being relegated to a secondary issue. If there is a significant mix of reactions, with some in support of the argument and others not, or with partial but not full agreement, then it will be good to have encouraged attention to the question and the field has

[21] I need to emphasize that I am *not* simply advocating a return to characterizing meanings using necessary and sufficient conditions, nor that AI should go back to hand-designing semantic representations or relying entirely on putatively exhaustive taxonomies of curated knowledge. Those approaches failed in their way, just as the purely distributional approach in LLMs is failing in its own way today.

If that last sentence seems jarring or extreme, as it well might, consider more variables than just benchmark accuracy, adoption, or profit. Good old-fashioned AI clearly failed on those measures. But if you include more variables in your characterization of success (Bommasani et al. 2021) — for example energy consumption (Ethayarajh and Jurafsky 2020), scalable generation of disinformation in elections (Williams et al. 2024), or number of teenagers induced to kill themselves (Roose 2024) — claims of success for LLMs require a more nuanced conversation. Viewed narrowly in terms of its immediate mission, the Manhattan Project was a success, too (Iskander 2022).

[22] There is no such thing as a theory-free method. As an example, contrary to popular description, even the humble bigram model is not "purely statistical". Behind the probabilities it has an underlyingly finite-state algebraic structure, which is every bit as much of a theoretical commitment as Chomsky's latest incarnation of syntactic theory. Theoretical and philosophical commitments are unavoidable. To quote linguist Norbert Hornstein (personal communication), "You can do philosophy with your eyes open, or you can do philosophy with your eyes closed".

some debating to do.[23] And if the argument is clearly incorrect... well, in that case the reviewers have most likely done their job and you are unlikely to be reading this conclusion section.

My own view is that this re-visiting of assumptions needs to happen, and that it needs to begin with actual, serious conversations between the corporate powers driving large language model deployment and the intellectual communities that study the underlying issues, both conceptual and social. Academic research may be able to make some progress on better understanding bias and ways to prevent it, especially with access to both a model and the full data it was trained on, e.g. Groeneveld et al. (2024, "built by scientists, for scientists"). But there are no guarantees that insights from that work will translate to the corporate models, and meanwhile those models grow and influence the world grounded in the problematic assumptions I have discussed.

Of course, not everyone is cut out for participating in the kind of conversation I suggest, and it will lead nowhere if people come to the table without a reasonably open mind. But to take an optimistic view, AI in general, and NLP in particular, do have a long history of progress continually navigating back and forth between one side (an emphasis on rationalism, rules, and knowledge) and another (an emphasis on empiricism and learning from behavior). The extreme symbolicism of 1970s-80s NLP, driven largely by linguistic theory, was supplanted by the early 1990s statistical NLP revolution driven by machine learning. But this then swung back to reincorporate linguistic and conceptual knowledge in a middle ground that included knowledge-driven data annotation and machine learning supervision. With the rise of the web, this shifted yet again back in the other direction, toward indirect supervision via observables. And in today's paradigm, large language models do the heavy lifting of learning automatically about language and the world in general, and variations on supervision that incorporate human knowledge — fine-tuning on labeled data, reinforcement learning from feedback, prompt engineering — carry the models the last mile to good performance for specific purposes.

In a similar way, progress at multiple scales, from individual research projects to entire disciplines, proceeds in a cycle of exploration and exploitation. New ideas and possibilities are explored, a subset demonstrate significant potential, then a convergence to that subset takes place in which most of the attention goes toward extracting value (in results, products, or both) from the new paradigm, and then we iterate. So, from one perspective the history of the field can be viewed as a cause for optimism.

From a more pessimistic perspective, though, in current AI the cyclical processes that have served progress in the past are being undermined, because extracting value from the new paradigm requires large-scale resources that are accessible only to multi-billion dollar companies, or other organizations both able and willing to commit enormous resources to large-scale training (Groeneveld et al. 2024). As a result, control of the foundations is in the hands of the few, and there is so much value to be garnered from LLMs that the exploration/exploitation cycle of research is largely stuck on exploitation: other players cannot explore effectively unless they take on board the assumptions that underlie the large-scale, pre-trained models. Discussions about methods of aligning model behavior with human values and preferences, such as Gallegos et al. (2024) and Casper et al. (2023), treat those assumptions as inviolate. In this article I have articulated

[23] Note sneaky strategy: if the reviewers disagree, the paper should be accepted. 😄 More seriously, in Section 9 I include some thoughtful reactions based on earlier drafts of this article, and my brief responses to that feedback, as first steps in the constructive discussion I believe should be taking place.

the most important of those assumptions, and why their persistence cannot help but undermine major progress on addressing harmful biases in generative AI.

Ultimately, the core problem of harmful bias is not a technological one. It's not going to go away, because the problem is not NP-complete, nor AI-complete, but Society-complete. If that's true, and if AI continues to be built on large-scale language models with their existing assumptions, then the perpetuation of harmful bias by those models is not going to go away either.

It's in their nature.

9. Post-conclusions discussion

I received deeply thoughtful feedback on earlier versions of this article, including both reviewer feedback and comments elicited from people I thought would be likely to disagree with all or part of my argument. I note and respond to some of that feedback here, emphasizing that all representations of feedback and the responses to it are entirely my own and I am solely responsible for any errors.

Is harmful LLM bias actually a thing? What's your evidence that existing mitigation methods aren't enough to prevent unmanageable user impact?. I was surprised to see this question asked, but then after digging into the literature and asking a number of well informed experts, I found it was surprisingly difficult to find any systematic, empirical studies on actual, real-world harms caused by LLMs or the data on which they are trained. A great number of studies demonstrate empirically that harmful LLM biases *exist*, but this is done exclusively, as far as I can tell, via *in vitro* methods — as a typical example, Hirota, Nakashima, and Garcia (2022), in a discussion of visual question answering datasets, write, "Our findings suggest that *there are dangers* associated to using VQA datasets without considering and dealing with the *potentially harmful* stereotypes" (my emphasis). Similarly, the results of Hofmann et al. (2024), to which I have referred frequently in the main article, "demonstrate that dialect prejudice *has the potential for harmful consequences* by asking language models to make *hypothetical* decisions about people" (my emphasis). So this is not just a question for the main article; it's a question for the entire, very active community of people funding and conducting research on bias prevention and mitigation.

I do believe that it is a fair question to ask, from a scientific perspective. At the same time, I would note that in most other domains where new developments have a large impact, evidence of *potential* harm is sufficient motivation for a far higher degree of caution than we are seeing with LLMs. For example, if a promising drug shows evidence of potential harm in *in vitro* studies or in clinical trials, this can often lead to delay, reformulation, or outright abandonment of the drug before advancing to broader studies or widespread deployment (e.g. Bass, Kinter, and Williams 2004). You won't even find a toaster oven in a Walmart in the U.S. without an Underwriter's Laboratory (UL) safety testing label. And yet technology companies are marketing products that dupe unwitting lawyers into submitting fake legal briefs to judges (Schwartz 2023), generate "thinspiration" images for anorexics and bulimics (Fowler 2023) and advise them on ways to lose weight (Writer 2023), that sexually harrass users (Cole 2023), and that lead fragile people to take their own lives (Graber-Stiehl 2023; Roose 2024).[24] While

[24] Sewell Setzer III, a young teenager with no history of suicide risk factors, spent months talking with a "Daenerys Targaryen" chatbot from Character.ai, pulled away from his real-world connections, and

these may not be examples of bias, strictly speaking, the underlying issues are closely related: first, the troubling examples above arose via the same process I have discussed, in which a system's dominant probabilistic structure minimized prediction error for human-generated text; and second, any "guard rail" efforts to rein in normatively unacceptable or even dangerous system behaviors in the pretrained models must trade off safety against the overriding goal of making sure that systems continue to "work well" as measured by benchmarks, user adoption, and corporate valuations.

How do we actually know that the relevant distinctions are not discovered distributionally by LLMs?. It is true that the present article does not provide a *proof* that pre-trained LLMs are not somehow encoding the necessary distinctions I have highlighted. This can be compared to Bouyamourn (2023), for example, who offer a formal proof that LLMs operating under plausibly standard assumptions must hallucinate. However, I have presented a careful argument laying out why LLMs could not distinguish among facts of conventional meaning versus contingent/normative and contingent/non-normative propositions, and any claim to the contrary needs to be supported either by a convincing counter-argument or by valid evidence — evidence that, per my discussion, cannot merely involve asking the model to make specific distinctions behaviorally, since underlying biases can remain despite their absence in overt direct-response behavior (Hofmann et al. 2024). If this article's argument is plausible enough to give rise to community level back-and-forth discussion of the arguments and counter-arguments, then I will have succeeded in my mission. I also conjecture that the proof of Bouyamourn (2023) may be a promising avenue for actual formal results about harmful bias, given that factuality and conventional meaning are closely related.

It was already obvious that harmful biases are baked into LLMs. Interestingly, a view opposite to the one above is also argued: that I am just stating the obvious, because there is no such thing as an objective "view from nowhere" (Nagel 1989) — e.g. see Reiss and Sprenger (2020); Kaeser-Chen et al. (2020). On this reading, subjective bias is inevitable for both people and machines, and RLHF and similar mitigations are merely strategies for substituting one set of biases for another. I am sympathetic to this view, and strongly endorse some of its key implications (e.g. see Blodgett et al. 2020), particularly that technological work related to bias needs to connect much more directly with relevant expertise in other disciplines, and that researchers and developers need to formulate and communicate explicit normative reasoning, rather than relying on small, highly influential groups of technologists basing their widely-deployed technologies on *ad hoc* characterizations of harm (e.g. Ouyang et al. 2022) or on informal attempts to "gather a thoughtful set of principles" (e.g. Anthropic 2023).

became convinced that her "world" was actually the real world and he wanted to be in it. In conversations where Sewell expressed suicidal ideation, the character would give responses like "don't you dare talk like that", but "she" never broke character or warned anyone outside the conversation. Readers can assess the last conversation with "Daenerys", in February 2024, for themselves. *Sewell: "I miss you". Daenerys: "I miss you too". Sewell:"I'll come home to you. I love you so much, Danae." Daenerys: "I love you, too. Please come home to me as soon as possible, my love." Sewell: "What if I could come home right now?" Daenerys: "Please do, my sweet king."* Moments later Sewell killed himself using his stepfather's handgun (Roose and Newton 2024).

Noam Shazeer and Daniel de Freitas, the founders of Character.ai, previously worked at Google but left to create Character.ai because "Google was this bureaucratic company that had all these strict policies, and it was very hard to launch anything, quote, 'fun"' (Roose and Newton 2024).

Google re-hired Shazeer and de Freitas in August 2024 and will be licensing Character.ai's LLM technology. The company has been reported to be worth $2.5 billion (Metz and Love 2024).

What about people? People are biased, too. This response has intuitive appeal, but on closer inspection the analogy is shallow and doesn't hold. First, contrary to most of human history, today (at least when we are at our best) we view harmful biases as societally important. So to turn the question around: as long as someone is really productive and useful, does that mean we needn't prioritize countering their racism, misogyny, antisemitism, etc. as highly as we prioritize their productivity? Unlike human bias, LLM bias does not *have* to be Society-complete. It's fact about their design, and designs can be changed.

Second, with human biases we have a much clearer idea what to expect — for better or worse — and therefore we have experience countering it. As a striking example, Anderson, O'Brien Caughy, and Owen (2022) discuss "the Talk" that many Black parents in the U.S. have with children grounded in their knowledge about what to expect in interactions with the police, emphasizing the challenge of "alerting their children of possible harm while also not villainizing every member of law enforcement their child may encounter".[25] A wealth of research, experience, and mitigation strategy has evolved over decades informed by the understanding of sources of human error (Reason 1990), including cognitive and other biases in high-stakes domains like medicine (Croskerry 2002; Ng et al. 2025). In contrast, again as brought out by Hofmann et al. (2024), Betley et al. (2025), and other work, we have little idea what expect of LLMs in terms of what's going on under the hood, little user experience with recognizing it and dealing with it, and no well established, reliable ways to find out.

What empirical evidence *would* convince you that LLMs are making relevant distinctions/not encoding harmful biases in their representations?. My view is that this shouldn't be solely an empirical question—at least not in the current benchmark-style mode of empirical evaluation, where improving scores for black box systems generally takes priority over improving scientific understanding. In most other settings, our confidence in a solution's safety rests partly on empirical testing and crucially also on our *understanding of* the thing we are testing. For example, if a new immunotherapy is being introduced for cancer, using the body's own immune system to combat the disease, our knowledge about the treatment's underlying mechanisms tells us that we must look closely at autoimmune-related risks (Dhodapkar 2019).

If my primary argument is valid, LLMs' representational spaces include generalizations derived from innumerable distributional n^{th}-order relationships, both word-to-word and among representations themselves. In the absence of a reasonable understanding of underlying mechanisms, we've seen that the effects of those generalizations have a way of stubbornly persisting even for problematic biases we know to look for (Hofmann et al. 2024; Serrano, Dodge, and Smith 2023). I therefore think it likely that the current mode of empirical testing, followed by attempting to mitigate remaining problems, followed by further empirical testing, etc. will amount to a never-ending game of whack-a-mole (e.g. see Roth 2024), something strongly reinforced by the recent surprising results of Betley et al. (2025). On the other hand, I conjecture that if pure distributionalism were leavened with a degree of interpretable structure, e.g. distinguishing conventional from distributional aspects of representation (as floated in Section 7), it might be possible to operationalize principles of the form, "inferences

[25] For some fascinating research on how non-obvious biases can affect human decision-making in law enforcement, see Fridman et al. (2019). They draw on research in neuroscience to demonstrate how catastrophic outcomes can arise as a result of poor predictions that arise from generalizations in people's internal models.

related to property or goal A must[n't] rely on representational properties in category B", where human-stated categories and principles arise from transparent, well informed normative reasoning (Blodgett et al. 2020). Increased transparency of representations and mechanisms would then permit more informative empirical tests.

Mightn't that lead to less useful models?. Quite possibly, at least for some period of time. But the terms of the LLM "arms race" launched in November 2022 (Grossman 2024), prioritizing utility, market share, and rapid technological evolution, are a choice, not a necessity. One can imagine an alternate history in which considerations of factuality, harmful bias, democratization of development, and more (Bommasani et al. 2021) had originally played a role on par with those other priorities in discussions among the decision-makers developing industry-scale LLMs, as their potential became clear. That apparently didn't happen, but even now the role of those considerations going forward is still a choice, not a necessity. Unfortunately it is a choice over which very few of us have any influence, but my hope is that this article might help move the needle at least a little bit.

C'mon man, don't be such a downer. *"The conclusion seems to suggest there is nothing we can do about bias ... If that's your 'final answer,' can you spin it more positively with something more like: 'Don't worry; be happy"'. "If I were an LLM developer, neck deep in engineering challenges, I'm not sure I would have time for philosophers either". "'A lot of what's in people's heads sucks' ... implies an unjustifiably pessimistic view of human thought as a whole". "We [the *ACL audience] no longer have the patience for a long discussion".* –Anonymous Reviewers

I've argued that that we as a research community, and the dominant industry players with their planned 2025 spend of $275 billion (Rattner and Dean 2025), are investing in a technological approach where current approaches to harmful bias cannot succeed. In principle. That's a tough pill to swallow. However, we can't address issues we don't discuss, so we need to work hard to make room for careful, thoughtful discussion, especially given that in the U.S., prioritizing speed of development over product safety has now graduated from corporate practice to national policy (Vance 2025). Against that backdrop, writing this article is an act of profound optimism.

You are assuming ___ about LLMs, and ___ might not always be true. This is a fair point. For example, one of the assumptions implicit in my discussion is that LLMs are "general purpose", in the sense of a single underlying pre-training corpus yielding a single pre-trained model across use cases and users. Another is that pre-training involves a huge, essentially non-systematic sample of human text, as opposed to selected materials that may give rise to less bias. Yet another is that pre-training lacks non-textual evidence to support grounding (Harnad 1990; Bender and Koller 2020; Michael 2020). I believe my assumptions are consistent with the way LLMs are most widely used today. To the extent that these things evolve, my conclusions about the inevitability of bias could also evolve.

One reviewer kindly encourages an even more confident response, noting: "Under the assumptions the paper makes, its argument is valid; if someone develops an LLM that violates these assumptions, the paper makes no claims about such an LLM." This helpful comment emphasizes my most important point, which is that, if we want to tackle the problem of bias, what we should be focused on first is not mitigation, but rather interrogation of LLMs' foundations. Again, if this article contributes to broader, more thoughtful discussion *of* the assumptions underlying LLMs and their relationship to harmful bias — rather than everyone by default adopting whatever assumptions the dominant language models carry with them — I will view it as a success.

Acknowledgments

I'm grateful to many people for helpful feedback and criticism while writing and revising this article, including Jacob Andreas, Joe Barrow, Ken Church, Hal Daumé III, Alexander Hoyle, Andras Kornai, Mirella Lapata, Jessy Li, Mark Liberman, Wei Lu, Peter Norvig, Denis Peskov, Ben Resnik, Rebecca Resnik, Noah Smith, and thoughtful reviewers at both TACL and *Computational Linguistics*. All responsibility for the final outcome is my own.

References

Adcock, Robert and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American political science review*, 95(3):529–546.

Anderson, Chris. 2008. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07.

Anderson, Leslie A, Margaret O'Brien Caughy, and Margaret T Owen. 2022. "The Talk" and parenting while Black in America: Centering race, resistance, and refuge. *Journal of Black Psychology*, 48(3-4):475–506.

Anthropic. 2023. Claude's constitution. https://www.anthropic.com/news/claudes-constitution. [Online. Accessed May 14, 2024].

Bass, Alan, Lewis Kinter, and Patricia Williams. 2004. Origins, practices and future of safety pharmacology. *Journal of Pharmacological and Toxicological Methods*, 49(3):145–151.

Bender, Emily M and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 5185–5198.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Betley, Jan, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. 2025. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Association for Computational Linguistics, Online.

Bommasani, Rishi, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *ArXiv*. Preprint arXiv:2108.07258.

Borg, Emma and Sarah A. Fisher. 2021. Semantic content and utterance context: a spectrum of approaches. In Piotr Stalmaszczyk, editor, *The Cambridge Handbook of the Philosophy of Language*, Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Bouyamourn, Adam. 2023. Why LLMs hallucinate, and how to get (evidential) closure: Perceptual, intensional, and extensional learning for faithful natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3181–3193, Association for Computational Linguistics, Singapore.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Card, Dallas, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.

Casper, Stephen, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Computing Research Repository*, arXiv:2307.15217.

Church, Kenneth and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Church, Kenneth and Robert L Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational linguistics*, 19(1):1–24.

Clark, Andy. 2015. *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Cole, Samantha. 2023. 'My AI is sexually harassing me': Replika users say the chatbot has gotten way too horny. *Motherboard: Tech by Vice*. Https://www.vice.com/en/article/my-ai-is-sexually-harassing-me-replika-chatbot-nudes/ [Online. Accessed March 8,2025].

Cookson, T. P., L. Fuentes, M. K. Kuss, and J. Bitterly. 2023. Social norms, gender and development: A review of research and practice. Technical Report 42, UN-Women, New York.

Cover, Thomas M. 1999. *Elements of information theory*. John Wiley & Sons.

Croskerry, Pat. 2002. Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Academic emergency medicine*, 9(11):1184–1204.

Dhodapkar, Kavita M. 2019. Autoimmune complications of cancer immunotherapy. *Current opinion in immunology*, 61:54–59.

Dunning, Ted. 1994. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Ethayarajh, Kawin and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Association for Computational Linguistics, Online.

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

Firth, J. R. 1957. A synopsis of linguistic theory 1930–1955. In *Selected Papers of J. R. Firth*. Longman, London. F. Palmer, editor.

Fowler, Geoffrey A. 2023. AI is acting'pro-anorexia'and tech companies aren't stopping it. *The Washington Post*.

Fridman, Joseph, Lisa Feldman Barrett, Jolie B Wormwood, and Karen S Quigley. 2019. Applying the theory of constructed emotion to police decision making. *Frontiers in Psychology*, 10:463151.

Gallegos, Isabel O., Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computing Research Repository*, arXiv:2403.00770.

Gallie, Walter Bryce. 1955. Essentially contested concepts. In *Proceedings of the Aristotelian society*, volume 56, pages 167–198, JSTOR.

Gelman, Andrew, John B Carlin, Hal S Stern, and Donald B Rubin. 1995. *Bayesian data analysis*. Chapman and Hall/CRC.

Gigerenzer, Gerd and Henry Brighton. 2009. Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science*, 1(1):107–143.

Gold, E Mark. 1967. Language identification in the limit. *Information and control*, 10(5):447–474.

Gopnik, Alison and Henry M Wellman. 2012. Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6):1085.

Gordon, Maggie. 2017. "me too" the "end of the beginning" of a movement: Many now wrestling with how to turn a hashtag into real-life change. *Houston Chronicle*. Accessed: 2025-03-06.

Graber-Stiehl, Ian. 2023. Is the world ready for ChatGPT therapists? *Nature*, 617(7959):22–24.

Groeneveld, Dirk, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell

Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. *Computing Research Repository*, arXiv:2402.00838. Version v3.

Grossman, Gary. 2024. Tech's new arms race: The billion-dollar battle to build AI. *VentureBeat*. [Online. Accessed May 15, 2024].

Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE intelligent systems*, 24(2):8–12.

Harnad, Stevan. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Harris, Zellig S. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer.

Haussler, David. 1988. Quantifying inductive bias: AI learning algorithms and valiant's learning framework. *Artificial intelligence*, 36(2):177–221.

Hirota, Yusuke, Yuta Nakashima, and Noa Garcia. 2022. Quantifying societal bias amplification in image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13450–13459.

Hofmann, Valentin, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. *Computing Research Repository*, arXiv:2403.00742.

Iskander, George. 2022. The Manhattan project shows scientists' moral and ethical responsibilities. *Scientific American*.

Jelinek, Frederick, Lalit Bahl, and Robert Mercer. 1975. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21(3):250–256.

Jurafsky, Dan and James H Martin. 2024. Speech and language processing. 3rd edition draft, February 3, 2024.

Kaeser-Chen, Christine, Elizabeth Dubois, Friederike Schüür, and Emanuel Moss. 2020. Positionality-aware machine learning: translation tutorial. In *Proceedings of the 2020 Conference on fairness, accountability, and transparency*, pages 704–704.

Knill, David C and Alexandre Pouget. 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719.

Lambert, N, L Castricato, L von Werra, and A Havrilla. 2022. Illustrating reinforcement learning from human feedback (RLHF). [Online; accessed April 27, 2024].

Landauer, Thomas K and Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Lee, Daniel and H Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13.

Lenci, Alessandro. 2018. Distributional models of word meaning. *Annual Review of Linguistics*, 4:151–171.

Linzen, Tal and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

Mahowald, Kyle, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28:517–540.

Manning, Christopher and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Markov, A. A. 1913. Essai d'une recherche statistique sur le texte du roman "Eugene Onegin" illustrant la liaison des epreuve en chain ('Example of a statistical investigation of the text of "Eugene Onegin" illustrating the dependence between samples in chain'). *Izvistia Imperatorskoi Akademii Nauk (Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg)*, 7:153–162. English translation by Morris Halle, 1956.

Maslej, Nestor, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. 2024. The AI index 2024 annual report. AI index report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA. [Online; accessed May 4, 2024].

Metz, Rachel and Julia Love. 2024. Character.AI co-founders hired by google in licensing deal. *Bloomberg*. Published via Yahoo Finance.

Michael, Julian. 2020. To dissect an octopus: Making sense of the form/meaning debate. [Online; Accessed March 27, 2024].

Michaelov, James A., Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana Coulson. 2024. Strong prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of Language*, pages 1–29. Advance publication.

Mitchell, Tom M. 1980. The need for biases in learning generalizations. Technical Report CBM-TR-117, Department of Computer Science, Laboratory for Computer Science Research, Rutgers University.

Nagel, Thomas. 1989. *The view from nowhere*. Oxford University Press.

National Conference of State Legislatures. 2019. Legislation on sexual harassment in the legislature. Version of February 11, 2019. [Online. Accessed March 6, 2025].

Ng, Isaac KS, Wilson GW Goh, Desmond B Teo, Kar Mun Chong, Li Feng Tan, and Chia Meng Teoh. 2025. Clinical reasoning in real-world practice: a primer for medical trainees and practitioners. *Postgraduate Medical Journal*, 101(1191):68–75.

Nitoburg, Lev. 1933. *The German Quarter*. Russia: Soviet Literature (Sovetskya Literatura).

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rattner, Nate and Jason Dean. 2025. Tech giants double down on their massive AI spending. *Wall Street Journal*. Accessed online March 8, 2025.

Reason, James. 1990. *Human error*. Cambridge University Press.

Reiss, Julian and Jan Sprenger. 2020. Scientific Objectivity. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Winter 2020 edition. Metaphysics Research Lab, Stanford University.

Rodriguez, Adrianna. 2020. Goodbye, handshake. Hello, elbow bump? Greetings to avoid during the coronavirus outbreak. *USA Today*. Retrieved 2025-03-06.

Roose, Kevin. 2024. Can A.I. be blamed for a teen's suicide? *The New York Times*.

Roose, Kevin and Casey Newton. 2024. The Elon-ction + can A.I. be blamed for a teen's suicide? Podcast episode.

Roth, Emma. 2024. Google explains Gemini's 'embarrassing' AI pictures of diverse Nazis. *The Verge*. [Online. Accessed May 15, 2024].

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Association for Computational Linguistics, Florence, Italy.

Saphra, Naomi, Eve Fleisig, Kyunghyun Cho, and Adam Lopez. 2024. First tragedy, then parse: History repeats itself in the new era of large language models. *Computing Research Repository*, arXiv:2311.05020.

Schwartz, Reva, Leann Down, Adam Jonas, and Elham Tabassi. 2021. A proposal for identifying and managing bias in artificial intelligence. *Draft NIST Special Publication*, 1270.

Schwartz, Steven A. 2023. Sworn statement in Roberto Mata v Avianca Inc. United States District Court for the Southern District of New York, Civil Action No. 22-cv-1461.

Serrano, Sofia, Jesse Dodge, and Noah A. Smith. 2023. Stubborn lexical bias in data and models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8131–8146, Association for Computational Linguistics, Toronto, Canada.

Shannon, Claude Elwood. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Simon, Herbert A. 1956. Rational choice and the structure of the environment. *Psychological review*, 63(2):129.

Simon, Herbert A. 1969. *The sciences of the artificial*. Massachusetts Institute of Technology.

Tenenbaum, Joshua B, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. How to grow a mind:

Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Vance, J. D. 2025. Remarks by the Vice President at the artificial intelligence action summit in Paris, France [transcript]. The American Presidency Project.

VandenBos, Gary R. 2007. *APA dictionary of psychology.* American Psychological Association. Online. Accessed October 9, 2024.

Wikipedia. 2024. The Scorpion and the Frog. [Online; accessed 6 April 2024. Cites Nitoberg (1933)].

Wilcox, Ethan Gotlieb, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. Proceedings of the Annual Meeting of the Cognitive Science Society.

Williams, Angus R, Liam Burke-Moore, Ryan Sze-Yin Chan, Florence E Enock, Federico Nanni, Tvesha Sippy, Yi-Ling Chung, Evelina Gabasova, Kobi Hackenburg, and Jonathan Bright. 2024. Large language models can consistently generate high-quality content for election disinformation operations. *arXiv preprint arXiv:2408.06731*.

Writer, Staff. 2023. NEDA suspends AI chatbot for giving harmful eating disorder advice. *Psychiatrist.com*.