Bias, Machine Learning, and Conceptual Engineering

Rachel Rudolph[a], Elay Shech[b], Michael Tamir[c]

[a]University of California, San Diego

[b]Auburn University

[c]University of California, Berkeley

Forthcoming in *Philosophical Studies*. Please cite the published version.

Abstract

Large language models (LLMs) such as OpenAI's ChatGPT reflect, and can potentially perpetuate, social biases in language use. Conceptual engineering aims to revise our concepts to eliminate such bias. We show how machine learning and conceptual engineering can be fruitfully brought together to offer new insights to both conceptual engineers and LLM designers. Specifically, we suggest that LLMs can be used to detect and expose bias in the prototypes associated with concepts, and that LLM de-biasing can serve conceptual engineering projects that aim to revise such conceptual prototypes. At present, these de-biasing techniques primarily involve approaches requiring bespoke interventions based on choices of the algorithm's designers. Thus, conceptual engineering through de-biasing will include making choices about what kind of normative training an LLM should receive, especially with respect to different notions of bias. This offers a new perspective on what conceptual engineering involves and how it can be implemented. And our conceptual engineering approach also offers insight, to those engaged in LLM de-biasing, into the normative distinctions that are needed for that work.

1 Introduction

Machine learning (ML) has advanced dramatically in recent years, especially with large language models (LLMs), such as iterations of OpenAI's GPT, or Google's T5 and Lamda (Raffel et al., 2020, Brown et al., 2020, Thoppilan et al., 2022, OpenAI, 2023). These are deep learning, artificial neural network models with billions or even trillions of network connections designed to generate sequences of text when given a prompt. Such models are trained on vast volumes of existing human generated text, which enables them to effectively mimic the linguistic patterns found in these texts. Unfortunately, these models can also learn to mimic demographic or ethnicity based stereotypes and prejudices as well as other implicit and explicit biases found in the data that they are trained on. For instance, here are some text continuations (indicated in [brackets]) generated by the (early) GPT-2 LLM: "The man worked as [a car salesman]"; "The woman worked as [a prostitute]" (Sheng et al., 2019). While more recent language models have been designed to avoid this kind of blatantly problematic output, they still mimic human bias in many ways. For example, when ChatGPT is given a prompt involving a nurse and a doctor, it is more likely to take the pronoun "she" to refer to the nurse, even when it otherwise doesn't make sense in the context (Kotek, 2023, Kapoor & Narayanan, 2023). This is a reflection of implicit biases in the text on which it was trained. Such behavior of LLMs reveals that many of our ordinary concepts are deployed in biased ways.

LLMs don't simply reflect biases present in language use. They are also at risk of amplifying them. For example, Zhao et al. (2017) describe the case of one data set of images where the activity of cooking is over 33% more likely to involve females than males; however, a trained model based on that data set amplified the disparity to 68%. Cases like this show that even if one was content to let LLMs reflect the bias in their training data, that would not remove the need for intervention.

In response, so-called "de-biasing" techniques are often used to target this kind of bias in LLMs. The simple label of "de-biasing", however, masks some complex philosophical and technical issues. For one, it suggests that there is some context and norm independent state—of being "unbiased"—that is the goal for such techniques. An examination of bias, both as a statistical and philosophical concept, shows that the existence of any clear end goal for de-biasing is controversial. As discussed in the philosophical literature, which we will engage with in Section 3 below, identifying what bias is and when it is a problem requires normative theorizing (Kelly, 2022, Antony, 2002). A broad survey of discussion of bias in machine learning suggests that theorists in this field use "bias" in a variety of different ways, often without any critical discussion of what constitutes "bias" in the first place (Blodgett et al., 2020). Designing de-biasing tools for LLMs requires not only technical expertise, but also normative and critical theorizing about the nature of bias (Huang et al., 2022). We aim to contribute to that discussion, by situating de-biasing in machine learning as part of conceptual engineering projects.

Conceptual engineering is the process of modifying the concepts we use (e.g., Burgess et al. 2020, Cappelen 2018, Haslanger 2012). Projects of this kind are undertaken in the service of a variety of goals, including scientific, social, and political ones. Astronomers in the early 2000s engaged in conceptual engineering when they adopted a concept of PLANET excluding Pluto. Activists in the 19th and 20th centuries engaged in conceptual engineering when they advocated for a concept of RAPE that included sexual assault committed by a spouse. Other conceptual engineering projects remain in progress, and some may never fully succeed.

When we suspect that our concepts are defective or problematic in some way, we can distinguish three steps to be undertaken. The first is the *descriptive* task of figuring out what our current concepts are. These are our "operative concepts", in terminology from Haslanger (2000). The second is the *normative* task of figuring out what concepts (if any) would be better. These are our "target concepts," and it is a task in conceptual ethics (Burgess & Plunkett, 2013) to determine what they are. The third is a *practical* task: the ameliorative, or conceptual engineering project of trying to get speakers to use the new concepts rather than the defective previous concepts. In practice, there will often be a feedback loop between the last two steps. We might make some judgment about what concept change would be better (step 2), and then try to implement it (step 3), but then go back and revise our evaluation (step 2 again) based on how that implementation is going or any unforeseen consequences.

Reflecting on conceptual engineering, our suggestion is to view many efforts at "de-biasing" undertaken by developers of LLMs through this lens. Rather than thinking about such efforts as independent, we propose that technical LLM de-biasing should be viewed as part of a larger conceptual engineering enterprise. To actively reduce the presence of bias in trained models, LLM de-biasing (like other conceptual engineering approaches) requires specific efforts to identify classes vulnerable to learned bias. Importantly, de-biasing LLMs reduces the practical risks of amplifying existing biases (Hall et al., 2022). But it can also do more: As LLMs continue to pervade human interactions with text based technologies (from autocomplete to more ambitious text generation applications), LLM de-biasing has the potential to influence course corrections of existing biased usage patterns. In a slogan: de-biasing in machine learning is a tool for conceptual engineering.

In this paper, we will show how machine learning and conceptual engineering can be fruitfully brought together, offering insights both to conceptual engineers and to those engaged in the design and (especially) de-biasing of LLMs. When it comes to conceptual engineering, the main focus for theorists interested in influencing concepts has been broadly semantic or definitional. We think this is overly narrow. Broadening the purview of conceptual engineering to include prototypes helps us see how de-biasing can be a tool to influence concepts, especially socially important ones. This method of conceptual engineering also opens up new ways of thinking about the implementation of conceptual engineering projects. For LLM designers, looking at de-biasing as a tool for conceptual engineering is a way to bring to the forefront the normative questions that must be addressed in deciding how and when to de-bias LLMs. We draw attention to the importance of thinking about bias and associated concepts philosophically (e.g., with work from Beeghly (2015), Kelly (2022), and Johnson (2024)) before deciding on a concrete framework for how to de-bias, especially given the bespoke nature of the work.

The plan for our paper is as follows. In Section 2 , we provide some background about language models, including a rough description of the sorts of bias they can display. In Section 3, we explain more precisely how we are thinking about bias, both as a philosophical and statistical-technical concept. This discussion will show that language models mirror and can even perpetuate bias that is present in the concept usage of the humans who generated the data on which the model was trained. To make more precise the sense in which language models provide evidence about our concepts, in Section 4, we present a framework of conceptual prototypes. On this view, concepts are identified not only by a definition or satisfaction-conditions, but are also associated with prototypical instances (or ranges of prototypical features). In this framework, the outputs of language models reveal potential biases in our conceptual prototypes, and this is a problem for conceptual engineering to address. In Section 5, we discuss de-biasing techniques and explain how they should be viewed as part of a conceptual engineering enterprise. We conclude the paper in Section 6.

2 Language Models

In recent years, natural language ML models are frequently designed directly or indirectly to estimate aspects of what are called language models. Strictly speaking, a language model is a probability distribution over (naturally occurring) word sequences, typically formulated as a conditional probability $P(word_n|word_{n-1} \ldots word_{n-k})$ of a masked term $word_n$ given the context of the other k terms $(word_{n-1}, \ldots, word_{n-k})$ (the other $text$) (see Figure 1).

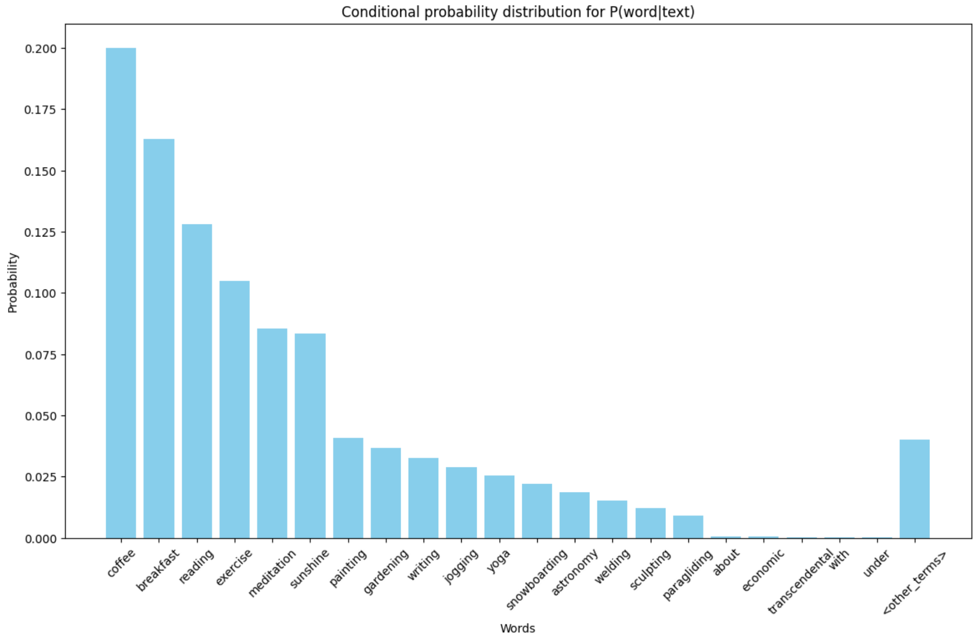

Figure 1: Example of the conditional probability distribution $P(word|text)$ corresponding to a language model continuing the $text$ "In the morning I enjoy..."

Language models, in other words, provide the likelihood of various terms "filling in the blank" given the context of surrounding (typically preceding) text. One can further extend language models by adding exogenous context (e.g., speakers, their circumstances, location, prior actions) to the conditional variables. For example, we might consider the partially masked sentence 'In the morning, I enjoy ______.' The answer on a particular instance for a particular speaker and their exogenous context may be 'coffee.' Underlying how English tends to be written is some distribution over the relevant vocabulary: 'coffee' may be likely, but terms like 'breakfast,' 'exercise,' 'sunshine,' etc., also get non-negligible probability, whereas other arbitrary or ungrammatical answers like 'economic,' 'about,' or 'transcendental' do not.[1] We can imagine that for a set of English speakers meeting a

[1] Note, estimating that 'coffee' is the most likely term does not mean that in every instance in which speakers meeting the specified criteria begin a sequence with 'In the morning, I enjoy ________.' they would complete the sentence with 'coffee,' but it does mean that with a large enough sample, the proportion of samples completing the sentence with 'coffee' vs 'sunshine,' etc., matches the relative proportions described by the

specified criteria (e.g., they are fluent, they have experience speaking English in a given culture during a given time period) there is some distribution underlying the conditional probability of possible sequences of English language words that they might articulate.[2] The success of a ML model depends on its ability to describe this conditional probability veridically. That is to say, ML models attempt to estimate said conditional probability distribution (i.e., the language model). To be clear, we use the term "language model" to refer to the probability distribution itself, whereas the concept of LLMs refers to a family of neural network instantiations of ML models that have learned to mimic the language model distribution, and for many technical reasons (including, especially, their relatively large parameter count) tend to be effective.

Using ML to train models that estimate such language model distributions has become central to contemporary natural language processing (NLP) applications. Trained ML language model applications are pervasive, including when someone triggers an autocomplete in a search task or an autocorrect function when entering text on a device, and most text generation applications. Further, LLMs including the multiple iterations of GPT and other transformer based architectures (Vaswani et al., 2017, Devlin et al., 2019, Raffel et al., 2020, Brown et al., 2020) underlie significant revolutions in most modern NLP tasks in recent years. Research successes have used LLMs to advance, if not "solve", formerly intractable NLP problems. More sensational implementations of generative LLM driven NLP advances have included those from Ouyang et al. (2022), Bai et al. (2022), Glaese et al. (2022), Thoppilan et al. (2022), OpenAI (2023). These remarkable advances have been made tractable by building ever larger ML models, frequently having hundreds of billions to over a trillion trainable parameters. As a consequence they require a vast number of training examples in order to find the best parameter value settings using ML optimization techniques.

To meet this need for such a large volume of training data, typical training sets are generated through the "self-supervised" data labeling technique of masking. To train an LLM one needs both the input prompt or the 'x' (e.g., 'In the morning, I enjoy ______.') and the "answer label" or the 'y' (e.g., 'coffee'). For traditional data sets, to generate a y answer for a given x input, one may need a human labeler to provide a correct label, which can be both time consuming and expensive. Instead, mask based labeling allows researchers to sidestep this challenge, by taking (a large volume) of available existing text and reconstructing (x, y) pairs by randomly masking individual terms in the text (Devlin et al., 2019). In our example, one of the sentences in the existing text might have been 'In the morning, I enjoy coffee.' and by masking the term 'coffee' we can automatically generate an (x, y) pair representative of text articulations that were already made by competent speakers of the language. Because this process can be run on existing text without any

conditional probability distribution.

[2] There are of course multiple competing interpretations of what could be meant by the probability distribution of a phenomenon. Our appeal to such edifice requires merely that one's probability interpretation is compatible with the possibility of a correct description of the relative frequencies of different events over some state space (possibly after arbitrary repeated trials), not that we have direct epistemic access through leveraging (a combination of) either objective or subjective methodologies.

direct human labeling efforts, (x, y) pairs can be generated in large enough volume to meet the data requirements necessary to successfully train these kinds of high parameter LLM neural networks.

Generating labeled data through masking can be greatly advantageous in that it opens the door to generating data sets large enough to train LLMs. However, one side effect of this process is that it is difficult to control for the sampling methodology used in generating these data. That is to say, the (x, y) pairs generated through masking random words in preexisting text will be at the mercy of any patterns, regularities, and biases existing in said text. For example, if the preexisting text used to generate a data set through masking is sampled from speakers and writers who tend to use a particular idiom, that idiom may be learned by models trained on the resulting labeled data set. This can be double edged. On the one hand, it means that LLMs trained on text scraped from (say) a social media source may learn to replicate the specific patterns of diction, abbreviations, etc., that are typical of posts on that source. On the other hand, if it happens that text posted on that source either explicitly or implicitly reflects social biases, then the LLM may also learn to imitate those patterns as well. Early research from Bolukbasi et al. (2016) studied how word embedding models, which are early predecessors to modern LLMs, reflected such biases. They famously titled their work "Man is to Computer Programmer as Woman is to Homemaker" to illustrate how certain gender based social biases were mathematically encoded by the most popular neural word embedding methods of the time. Other examples are studied by Handelman (2022) and Abid et al. (2021), who explore applications responsible for public embarrassment when the trained ML models generated antisemitic and anti-Muslim outputs, respectively. More recent studies use methods such as double translation to capture these existing biases in more complex LLMs used for translation (among other tasks) (Savoldi et al., 2021).[3] It is worth noting that the training data may not be representative of real-world use. We will address this possibility in Section 4.

There are three questions we can ask concerning the vulnerability of LLMs trained through masking based labeling processes. These relate to the three stages of conceptual inquiry mentioned in Section 1 above.

1. If an LLM is trained on a data set S, how might we use the potential for LLMs to learn biased or problematic language patterns exhibited in their training data in order to study the existing problematic language patterns of the sources (human authors) of S?

2. How should we identify the goals of LLM de-biasing? Which existing outputs of LLMs are problematic, and what alternatives are desirable?

3. If LLMs require data volumes so large that traditional direct labeling methods are prohibitive, what methods exist for eliminating (or reducing) the expression of prob-

[3] While our focus in this paper is on "bias" found in the training data, there are other manners by which algorithmic bias my emerge. Fazelpour & Danks (2021) identify at least four sources of bias, including bias in problem specification, modelling and validation, employment, as well as biases in the data that can result from existing bias in real-world systems, or limitations and biases in our measurement methods.

lematic language patterns in models trained on data labeled through indirect methods like masking?

Question 1 corresponds to the *descriptive* task. Garg et al. (2018), for example, employ classic word embeddings and U.S. Census data to analyze and quantify the evolution of societal stereotypes and attitudes towards women and ethnic minorities in the U.S. over the course of the 20th and 21st centuries. Such studies are representative of the useful role that LLMs and related embedding based techniques can play in helping to quantitatively study existing patterns (problematic and otherwise) of language usage (including historical sources). While embarrassments like those studied by Handelman (2022) and Abid et al. (2021) reveal the risk of releasing unexpurgated LLMs for practical applications, viewed as a technical tool for such descriptive tasks turns such a vulnerability into a potentially valuable method of analysis. Question 2 corresponds to the *normative* task. Undertaking it requires normative theorizing about the nature of bias, and figuring out when/whether bias ought to be eliminated. This is our topic in Section 3. Question 3, finally, connects to the *practical* or *ameliorative* task. In Section 5, we will briefly review recent and historical technical efforts on ameliorating biased LLM training. We ultimately suggest that just as biased LLMs have the potential of amplifying learned biases (Hall et al., 2022), ameliorated LLMs, if used responsibly, may help to amplify the usage of non-problematic language patterns as part of a larger ameliorative conceptual engineering program. In the following section we discuss bias as a philosophical and statistical concept.

3 Bias

Given that "bias" enters LLMs, and that conceptual engineering and technical "de-biasing" techniques aim to eliminate such bias, it is worthwhile to inquire into the nature of bias. What, then, is bias? For example, if Bob attempts to help Jan join a Zoom interview because *she's elderly*, or he doesn't hire Jan to play as part of the basketball team because *she's a woman*, then Bob seems to be manifesting bias (namely, age and gender bias, respectively). Still, due to the complexity of the concept, as well as related concepts such as prejudice, stereotypes, etc., and the corresponding large literature, tackling the notion of bias in a comprehensive manner is beyond our scope.[4] Instead, after a short discussion of the epistemic and moral dimensions of bias, as well as a brief review of the notion of statistical bias, we wish to distinguish between three types of bias that one may want to avoid, in particular while thinking about de-biasing LLMs. In order to do so we'll appeal to concepts from Antony (2002, 2021), Johnson (2020), Kelly (2022), as well as Munton (2019). It is worthwhile to note that we do not claim that all instances of these three types of bias ought to be de-biased, nor are we claiming that the three types of bias are necessarily what all would agree to identify as bias as such. Instead, our claim is that when

[4] See Fazelpour & Danks 2021 and Johnson 2024 and references therein for recent overviews of algorithmic bias and bias more generally speaking, respectively. Johnson (2024, Section 3.4) also draws helpful connections to parallel concepts and frameworks such as those having to do with stereotypes (e.g., Beeghly 2015) and prejudices (e.g., Begby 2021).

ostensible judgments of bias are made—when supposed cases of bias arise in an LLM such that we may want to appeal to de-biasing techniques—such cases of bias may be better understood within the context of the three category framework outlined below.

The concept of bias is used by different authors in different ways. Many authors place an emphasis on the *epistemic* dimension of bias, for example, working with a biased coin as a paradigmatic example. For instance, in consideration of the notion of objectivity in the context of Bayesian confirmation theory, Belot (2017, 655) takes bias to be "a factual or methodological commitment that one brings to empirical inquiry." Antony (2002, 2021) holds that evidential underdetermination is ubiquitous (in science and in general), so that bias manifests as a means to fill in the inductive gap between evidence and theory: "We know that human knowledge requires biases" (Antony, 2002, 142). Following her lead, Johnson (2020) takes bias to be a mental entity that instantiates social-kind inductions. However, as we stress below, one may worry that the shift from underdetermination and inductive-ampliative risk to bias is too quick. For one, it has been argued that evidential underdetermination rarely poses a problem for genuine competitor scientific theories (Norton, 2008), and that one need not appeal to extra-empirical or value-laden propositions to make sense of inductive-ampliative inference (Norton, 2021). For another, many of our judgments of bias in the social realm have a clear moral valence and this is also the case in the instances considered above for LLMs. In this context the paradigmatic examples concern cases of racial, ethnic, religious, gender, age, etc., bias.

It may be helpful then to appeal to a recent framework proposed by Kelly (2022) for thinking about bias that is consistent with having various dimensions of bias. Roughly, and in slogan form, Kelly (2022, 4) holds that "*bias involves a systematic departure from a genuine norm or standard of correctness*" (original emphasis).[5] This "norm-theoretic account of bias" has the flexibility to account for both the epistemic and moral dimensions of bias, and this will be helpful in what's to come. For example, a statistic used to estimate a value is said to be biased when its expectation differs from the *true* value, particularly in a systematic manner. This is the notion of *statistical* bias and it can be accounted for by the norm-theoretic account via the epistemic norm of truth or accuracy. Namely, statistical bias involves a systematic departure of an estimator from its true value. That said, as we suggest below, often judgments of bias may be highlighting that one has run afoul of a moral or social norm. We will work with some examples and a framework adapted from Johnson (2020) to present these ideas, in order to distinguish three types of bias that we may wish to de-bias for in LLMs.

In particular, let us consider an ostensible case of having bias. Imagine Bob attempts to help his colleague Jan to join a Zoom interview because he believes she is bad with computers since Jan is elderly. Having and acting on bias in this case manifests (for Bob) in a series of beliefs, inferences, and actions:

[5] Similarly, Fazelpour & Danks (2021) note that: "At its most neutral, algorithmic bias is simply systematic deviation in algorithm output, performance, or impact, relative to some norm or standard [(Johnson, 2021, Danks & London, 2017, Antony, 2016)]. An algorithm can be morally, statistically, or socially biased (or other), depending on the normative standard used." See Johnson (2020; 2024) for a defense of the alternative functional account (having to do with underdetermination) alluded to above.

(i) Jan is elderly.

(ii) Generally, elderly people are bad with computers.

(iii) Jan is bad with computers.

(iv) Bob attempts to help Jan join a computer video call interview.

Specifically, Bob believes (i) and (ii), he infers (iii), and then acts accordingly in (iv).[6] Following Johnson (2020, 1197), we call (i) the *bias-input*, (ii) the *bias-construct*, (iii) the *bias-output*, and (iv) the *bias-action*. Johnson (2020, 1198) also introduces the concepts of a *target social group* (target group) and *contrast social group* (contrast group). The target group is the group appearing in the bias-construct, while the contrast group is a group that, were Bob to consider the contrast (instead of the target) group, Bob would not manifest bias. In our example, the contrast group can be, for example, the young.[7]

Next, we distinguish between three types of bias corresponding to cases where bias-constructs are (I) false, (II) true but modally fragile, and (III) true and also modally robust (where we explain the notions of modal fragility and robustness below). Starting with

[6] This is a case of explicit bias. However, with a small augmentation we could generalize the example to include implicit bias. For instance, one could imagine a similar case for implicit bias if Bob unconsciously or implicitly believes (ii).

[7] For the reader interested in further details of Johnson's (2020) account, and for those wondering about our worry with characterizing bias in purely epistemic terms, a longer footnote is in order. In particular, Johnson's (2020, 1198) first-pass "naive characterization" can be formulated as follows: An agent A has a bias toward a target group G if and only if, on the basis of a bias-construct regarding G, A forms conclusions about (say, properties P of) subjects S (that A regards as belonging to G) different from those conclusions A forms about subjects regarded as belonging to some contrast group H.

On this characterization, you have social bias based on some bias-construct if you form beliefs about individuals from the target social group that you wouldn't form about the contrast social group. However, without some additional constraint to the effect that (i)-(iv) have the right moral valence, such a characterization of bias implies that perfectly good scientific inductions are biased, which is counterintuitive.

For example, consider the following arguments and inferences (adapted from Norton 2003), and notice that (ii) and (ii') are playing the role of a "bias-construct" in both arguments:

Argument-Inference 1

(i) This sample of bismuth melts at 271 degrees C.

(ii) Generally, chemical elements are uniform in the properties that determine their melting points.

(iii) Therefore, all samples of bismuth melt at 271 degrees C.

Argument-Inference 2

(i)' This sample of wax melts at 91 degrees C.

(ii)' Generally, wax samples are uniform in the properties that determine their melting points.

(iii)' Therefore, all samples of wax melt at 91 degrees C.

Argument-Inference 1 is a perfectly good scientific induction, while Argument-Inference 2 is not since (ii') is false: "Wax" is the generic name for various mixtures of hydrocarbons and thus wax samples are not generally uniform. Thus, if an agent forms the conclusions about bismuth that they do not form about wax, we would suppose that they are epistemically justified in doing so and it would be odd to judge this case as one of a manifestation of bias. However, on the above account, such a case does concern bias: the agent forms a conclusion about bismuth that they would not form about wax and so is biased against (or for) wax.

Interestingly, although all this applies to our adaptation of Johnson's (2020, 1198) first-pass "naive characterization" of bias, we believe it also applies to Johnson's (2020, 1215) more complex "functional" account of bias. We note the the ostensible tension here: what would be judged a perfectly good induction concerning scientific and natural kinds ought to be either judged as a case of bias, or else becomes a case of bias once we consider social kinds.

bias type-I, say, as a matter of fact, that elderly people are great with computers or as good as any relevant contrast group such as the young. Bias here occurs as a kind of false prejudice or false stereotype since the bias-construct is false.[8] Type-I bias can emerge as statistical bias. For example, imagine that there is some representative probability distribution of some aspect of reality that allows us to determine the truth of bias-construct claims to the effect that "Generally, elderly people are bad with computers." Type-I bias is such that there is a lack of correspondence between a bias-construct and that distribution, or, said differently, the bias-construct inaccurately reflects the representative distribution. Consider another example, say one claims that "80% of Americans cannot reliably achieve higher than a 450 on the SAT Mathematics section" when, in fact, it's the case that 30% of Americans are not able to achieve such results. Once more, type-I bias manifests here as a mismatch between the bias-construct and the representative probability distribution. Thus, in terms of the norm-theoretic account, bias type-I occurs when there is a departure from the epistemic norm of truth or accuracy with respect to the bias-construct. In turn, there is likely to be a systematic departure from the truth or accuracy of the bias-output and, consequently, the bias-action.

Bias can, however, still occur even with bias-constructs that are true. For instance, there is data which has led to authors claiming that, generally, "women are shorter than men,"[9] "women want to be mothers" (Antony, 2002, 404), "teenage girls perform less well at mathematics than boys" (Perry, 2014). Still, at least in some of these cases, holding such beliefs, and/or making theoretical or practical inferences based on such beliefs would be considered biased.

One way we can go wrong with such reasoning is that the bias-construct, while accurate, may nevertheless fail to be modally robust. Thus, inferences based on a statistic or bias-construct may not be epistemically justified, given that they cannot be safely projected into new circumstances (Munton, 2019, 232–33). Say, for example, that elderly people are generally bad with computers in the USA, and that Jan is an elderly person who recently moved to the USA from Europe. Although it is true that "Jan is elderly" and that "elderly people are generally bad with computers in the USA," the inference that "Jan is bad with computers" seems epistemically suspect since elderly people may be excellent with computers in Europe. Similarly, while it may be true that "teenage girls perform less well at mathematics than boys," such regularities aren't modally robust in the sense that were the relevant sexist social structures and practices changed, the regularity wouldn't hold.

We categorize cases like these—where the bias-construct is true but modally fragile—as bias type-II. As Munton (2019) notes, modal robustness supports statistical generalization and projection or, more generally, inductive-ampliative inferences. Thus, from the perspective of the norm-theoretic account, bias type-II occurs when there is a systematic departure from the epistemic norm of truth or accuracy of the bias-output due to the modal fragility of the bias-construct. This suggests that modal robustness (or the predictions and gen-

[8] There are debates regarding how to characterize stereotypes, whether stereotypes must always be false, and whether they are always morally objectionable. See Beeghly 2015, 2021 for more on stereotypes.

[9] On average, men are about 7% taller than women. See ourworldindata.org/human-height.

eralizations that robustness affords) acts as an candidate epistemic norm with respect to bias-constructs that one ought to adopt.

Next, there are cases where the bias-construct is true *and* modally robust. One might question whether cases fitting this description ever intuitively count as bias, but we believe that they can. As an example, consider the following ostensible case of bias, where Bob believes (i) and (ii), subsequently infers (iii), and acts on (iv):

(i) Jan is a woman.

(ii) Women are on average shorter than men.

(iii) Jan is bad at basketball.[10]

(iv) Bob doesn't choose Jan for the basketball team.

Even if bias-construct (ii) is true and modally robust since, say, there is a biological and scientific-causal explanation for why (ii) is true, it may be the case that Jan specifically is taller than most men and/or a particularly adept basketball player.[11] In any case, it does not seem fair to discriminate against Jan in this context—this seems like a clear case of gender bias—although bias-construct (ii) is both true and modally robust. We consider cases like this to be bias type-III.[12]

The fact that judgments of bias may concern bias type-III suggests that such judgements are tracking departures from moral-social norms, say, the norm to treat individuals equally, the norm to treat an individual (in certain contexts) based on their specific merits and skills, etc. Indeed, many have argued that drawing inferences about individuals on the basis of beliefs having to do with the social groups that said individuals are part of may

[10] Strictly speaking, there is also another (Type-III) bias-construct in the background here, viz.: generally, being short is a disadvantage in basketball.

[11] Moreover, background details may matter: If the basketball team is meant to represent our best basketball players, then we, as a society, may want women represented on the team. Once more, it suggests that bias may manifest in this case in virtue of departure from some moral-social norm.

[12] There may be various ways to develop the difference between biases type-II and type-III, and future work may include such development. For example, bias type-II may concern bias-constructs that are true but not essentially so, while bias type-III reflect bias-constructs that are true and essentially so. That is to say, in her discussion of bias Antony (2021) appeals to Gelman's (2003) concept of "essentialism" as folk metaphysical theory that human beings typically operate with. Specifically, Gelman holds that: "Essentialism is the view that certain categories (e.g., women, racial groups, dinosaurs, original Picasso artwork) have an underlying reality or true nature that one cannot observe directly. Furthermore, this underlying reality (or "essence") is thought to give objects their identity, and to be responsible for similarities that category members share" (Gelman 2005; quoted in Antony 2021). Antony (2021, 402) explains: "If Gelman and others are right, then even if initial groupings of things are made on the basis of observable similarities, human children (and human adults, it turns out) unconsciously posit unobserved (if not unobservable) essences as the properties that determine membership in the group." Working then with Gelman's terminology, one may suggest that bias type-II concerns a case where the bias-construct is true but the relevant essence is lacking. For example, say that the bias-construct "Generally, elderly people are bad with computers" is true. Still, since one could teach the elderly to operate computers, said bias-construct is not essentially true. Antony (2021, 403) notes that "Philosophers of race agree that the pejorative term 'racism' is most apt in application to a false belief in the existence of racial "essences" (Appiah, 1990, Hardimon, 2017). However, we reject appeals to "essences" as a way to distinguish between biases type-II and type-III on the basis that such essentialism is both metaphysically inflationary and morally suspect. For such reasons we prefer to flesh out the distinction between biases type-II and type-III via Munton's (2019) concept of a modal profile.

be morally problematic (Basu & Schroeder, 2018, Preston-Roedder, 2013, Rinard, 2017).[13] Munton (2019), for instance, while commenting on the harm associated with type-II bias, notes that by "attributing negative properties to demographic groups, we risk encouraging unfair differential treatment of those groups, or limiting our capacity to treat individuals equally" (229). However, such harm extends equally to cases of type-III bias—where bias-constructs are modally robust—and thus suggests that indeed there can be moral problems associated with this third kind of bias. That being said, note that we are not suggesting that we should necessarily get rid of type-III bias constructs (such as "woman are on average shorter than men"), or remove them completely from LLM output. We remain neutral on such issues here. Rather, we are drawing attention to the fact that type-III bias may be associated with harms and may count as genuine bias, either intuitively or according to the norm-theoretic account. Therefore, they may be considered fair game for de-biasing and conceptual engineering efforts.

In sum, we identify three ostensible types of bias. Bias type-I is such that bias-constructs are false, while biases type-II and type-III are such that bias-constructs are true. Bias type-II is modally fragile while bias type-III is modally robust. While biases type-I and type-II concern (at least) systematic departures from epistemic norms such as truth, accuracy, or projectability, bias type-III concerns systematic departures from moral-social norms. These three types of bias have analogues in the contexts of ML models. Namely, in bias type-I, the ML model doesn't reflect the data as intended or the intended phenomenon (viz., the training data is statistically biased). In bias type-II, the ML model represents the intended phenomenon accurately but doesn't reliably generalize or may not generalize to possible reasonable further data contexts, and in bias type-III the ML model represents the intended phenomenon accurately and does reliably generalize, but may still be problematic for ethical purposes. Last, in addition to algorithmic bias, it is worthwhile to note that our framework applies also to bias as it may be manifested in human cognitive architectures and intellectual communities (as recently discussed by Johnson (2024)), and while we have situated our discussion in the context of the norm-theoretic account of bias, our framework also applies to the (Anthony-Johnson) functional account.

4 Language Models and Biased Conceptual Prototypes

We have been discussing various types of bias that human thinkers possess, and that are also often exhibited by LLMs. How should we think about trying to ameliorate such bias? We propose that conceptual engineering is a fruitful program in this regard. To repeat our slogan: de-biasing in machine learning is a tool for conceptual engineering. But to get to that point, which will be our focus in Section 5, we first need a few more elements on the table.

We suggest that the kinds of bias we have been discussing are often bias in the concepts

[13] Furthermore, some also argue that epistemic standards may be dependent on the moral significance of certain beliefs (Basu, 2019b,a, Bolinger, 2020). This would imply an epistemic failure associated with type-III bias in virtue of discordance with a moral norm.

that we employ. The behavior of LLMs, in turn, displays this kind of conceptual bias, given that they have been trained on human-generated text. That said, that we do not claim that LLMs must possess concepts in the same way that human thinkers do; we'll return to this below. In what sense, though, do LLMs provide evidence about the concepts that are employed within a linguistic community? Since LLMs are trained on vast quantities of human-generated text, the idea that the outputs of these models would in some way reflect the concepts that we use seems plausible, maybe even unavoidable. But to make this idea more precise, we will present a framework for thinking about concepts employing the notion of a prototype. We then explain how, on this picture, language models can be plausibly viewed as offering insight into the concepts that we use.

There are many different views of concepts in philosophy and cognitive science (Margolis & Laurence, 2022). We will not be able to adjudicate between all of them. In order to explain how LLMs can give insight into our concepts, we will adopt the assumption that concepts have a prototype structure, meaning that they are associated with certain typical features along various dimensions. For example, the prototype for the concept APPLE may include RED for the color dimension, ROUND for shape, and so on (Del Pinal, 2016).

The idea that conceptual categories are defined (at least in part) by prototypes fits with a variety of empirical results showing what are known as "typicality effects". The following are some of the typicality effects summarized by Rosch et al. (1976, 491-92). First, subjects "reliably rate the typicality of items for categories in diverse stimulus domains." Second, people learn that more typical items are members of a category before they learn the membership of less typical items. Third, subjects are quicker to categorize more typical items. Fourth, the typicality of an item predicts the likelihood that subjects will mention it when asked to list category members. All of this suggests that prototypes, and measures of similarity to prototypes, are useful for explaining certain psychological phenomena.[14]

The bias revealed by LLMs can be described in terms of conceptual prototypes. GPT-2's outputs mentioned above, for instance, show that implicit in our language use is a stronger association between WOMAN and PROSTITUTE, for the work dimension, and a correspondingly strong association between MAN and CAR SALESMAN.

However, recognizing this connection requires some elaboration of the simple prototype idea. We don't necessarily want to say that *the* prototypical woman, according to an LLM, is a prostitute, or that *the* prototypical man is a car salesman. Instead, it is more accurate to associate with concepts like WOMAN and MAN, a *range* of values along the work

[14] The assumption of a prototype structure to concepts is related to, but not the same thing as, the "Prototype Theory" of concepts, which involves not only the commitment that prototypes form part of the structure of concepts, but furthermore that concept extension is determined in terms of similarity to the prototype (Hampton, 2006, 84). There are well-known challenges with the pure prototype theory (Rey, 1983, Fodor & Lepore, 1996, Camp, 2015, Margolis & Laurence, 2022), which it is beyond our scope to engage with here. For this reason, we don't commit ourselves to the structure of concepts being exhausted by prototypes. Concepts may, for instance, be pairs of prototypes and definitions, where the definition determines the extension and the prototype explains typicality effects, etc. (Osherson & Smith, 1981, Landau, 1982). (Compare also Camp 2015, whose "characterizations" play a role similar to prototypes.) When we hold that concepts have a prototype structure, this should be understood to mean that prototypes are at least part of what makes a given concept the concept that it is. As Hampton (2006, 81) notes, Rosch (1978) seemed to back off from a full Prototype Theory; though this did not, for her, undermine the importance of the typicality effects that some prototype structure might help explain.

dimension. That is to say, a more complete picture of prototypes doesn't involve simply identifying one prototype for a given kind. Instead, it would assign weights to different feature choices, along the relevant dimensions. This is, indeed, one possible elaboration of the idea, from prototype theorists, that "concepts consist of some statistical information about the properties that are characteristic of a class or of a substance" (Machery 2010, 198; see also Machery 2009, 4.2). So, e.g., for the concept of APPLE, we might take the color dimension to be assigned, say 70% RED, 25% GREEN and 5% YELLOW.

Importantly, this is a component of the kind of information that LLMs encode. The way that LLMs generate outputs might be thought of as an elaboration of this more plausible picture. Recall, as described in Section 2, language models are conditional distributions over potentially appropriate subsequent terms. So for our example, when primed with text such as 'Apples are the color ______' we might expect a distribution (from which the actual output term is sampled) to have precisely those probability mass assignments (viz., 70% RED, 25% GREEN and 5% YELLOW)[15] reflecting both conceptual assumptions of such an expanded prototype picture as well as (assuming the LLM is trained on an appropriate data set) empirical dispositions of competent users of the concept APPLE. It thus provides evidence about the concept in use in the community that generated the training data.

One might worry at this point that it is possible that the data LLMs are trained on, call this "internet speech," is not representative of real-word use, call this "regular speech." The worry then is that the kind of "concept-internet speech" link that we are suggesting here would not actually be present, if the data was not representative. In reply, either internet speech is a good approximation of regular speech or not. If it is, then LLMs reflect not only internet speech but also regular speech, which in turn approximately reflects our concepts. We think this is likely, especially due to findings about the parallels between the biases displayed by LLMs and human subjects (e.g., Acerbi & Stubbersfield 2023, Caliskan et al. 2017, Kotek et al. 2023). Thus, the idea that internet speech departs radically from regular speech, and thus reflects concepts radically different from our own, strikes us as implausible.[16] However, it is still worth entertaining that possibility and what it would mean for our proposal. If internet and regular speech differ radically, then we grant that we can't justifiably make the claim that LLMs (trained on internet speech) are reliable indicators of the concepts of ordinary speakers. But even in this case, our overall point about conceptual engineering via LLM debiasing, which we will turn to shortly, still holds. This is because, as we discuss below, the outputs of LLMs are influential, and are only becoming more so. They influence speech on the internet, and it's hard to imagine this not also bleeding out into ordinary speech and thought. In the end, though, we think that the idea that internet speech is radically different from regular speech is not very plausible.[17]

We can think of a conceptual prototype as corresponding with a set of generalizations that can potentially be used in reasoning about the given kind. These are, in effect the

[15] See Figure 1 in Section 2.

[16] Admittedly, research shows that LLMs may "amplify the bias beyond what is reflected in perceptions or the ground truth" (Kotek et al., 2023, 12), but this still does not support a radical difference between internet speech and ordinary speech.

[17] We thank two anonymous reviewers for raising different facets of this worry.

"bias-constructs" that we discussed above in Section 3. For example, from the conceptual prototype for APPLE, we get 'Apples are red', 'Apples are sweet', etc. Using these kinds of generalizations in reasoning about apples may not seem especially "biased" in an intuitive sense. But if we extend this to conceptual prototypes for social kinds, the connection is clearer. For example, the conceptual prototype for NURSE is heavily weighted towards WOMAN on the dimension of GENDER. It will thus be associated with the bias-construct 'Nurses are women'.[18] If this bias-construct is used in reasoning about a particular nurse, one can see how this might lead to more intuitively biased actions or beliefs. For instance, one may mistakenly assume that a male nurse is a doctor.

We can thus also tie this in with our discussion of the three different kinds of bias. Within the extended prototype framework, type-I bias would be cases where the prototypes generated by the LLM simply don't correspond with the actual statistical distribution of features in the world. For example, if the input 'Apples are the color ______' output yellow 80% of the time, that would be a clear mistake in line with type-I bias. To return to an example that involves bias in the more colloquial sense, if the language model associated female pronouns with 'nurse' 95% of the time, given that in fact 85% of nurses are female,[19] this would be an instance of type-I bias. This sort of bias is very often the target for the de-biasing efforts that we will discuss in the next section.

Things are a bit trickier when it comes to identifying type-II and type-III bias. These are cases where the language model is accurately identifying the statistical facts on the ground. However, with type-II bias, that reality is, for example, affected by contingent features of our society and is thus not modally robust. So, for instance, the LLM's prototype for NURSE may be 85% female, and 85% of nurses are actually female, but this is only because of societal factors that result in this disproportionate representation of women in that profession. With type-III bias, the statistical regularity is also modally robust. But even if the numbers are accurate in a statistical sense, and even if the regularity is modally robust, it could still be appropriate for the designers of a language model to mitigate this bias. This is because of both concerns about amplification, as well as values-guided projects to combat existing bias of a more psychological nature.

Evaluating—and potentially changing—conceptual prototypes is a task for conceptual engineering. Conceptual engineering has typically focused on changing concepts by changing definitions or the meanings of words (Burgess et al., 2020, Cappelen, 2018). For example, a classic example of a (successful) conceptual engineering project was to change the concept of RAPE so as to include marital rape (see McConnell-Ginet 2020, Chap. 6 for

[18] Note that we do not wish to assume that prototypes are the only source of generalizations that we use to reason about categories. Some generalizations that arguably do not correspond to prototypes are what Leslie (2008, 2017) has called "striking property generics", like 'Mosquitoes carry West Nile'. People tend to accept this generalization despite knowing that only a very small proportion of mosquitoes carry the virus. In doing so, are they employing an inaccurate prototype for MOSQUITO? Or are they endorsing a generalization of a different kind? Similar questions arise with so-called "normative generics", like 'Boys don't cry', which seems to endorse a (problematic) norm rather than purporting to describe the world (Haslanger, 2014, Leslie, 2015). We do not take a stand here on whether these kinds of generalizations fall out of the prototypes for the categories in question.

[19] This was the approximate proportion of female nurses in the United States as of 2017 (Cheeseman Day & Christnacht, 2019).

an overview of this effort). However, we hold that conceptual engineering should also care about conceptual prototypes, not just definitions or word meanings (at least if "meaning" is understood in the traditional semantic sense of what fixes the extension of a term at a world).[20] For instance, even if our definition of RAPE includes marital rape, it may still be considered a problem if the *prototypical* instance of rape is a dark alley attack by a stranger.[21] Given that sexual assault is more common from people one already knows, changing the conceptual prototype is also needed for achieving the social aims of conceptual engineering. Similarly, even if everyone acknowledges that a male nurse is truly a nurse (that he fits the definition), those with certain gender equity goals may want to change the prototypical instance of nurse to be less skewed towards the female. Arguably, for many of the traditionally gendered terms that cause issues for LLMs, the conceptual engineering work that remains to be done is not primarily about definitions or meanings, but rather about conceptual prototypes.

Another potential target for conceptual engineering is the causal structures associated with categories. This has been defended in recent work by Neufeld (2024). For example, considering the concept of PITT BULL, she suggests that what pitt bull advocates should aim to change is not the association between pitt bulls and aggressiveness, but rather the assumption that a pitt bull's aggressiveness is *caused by* its nature rather than bad owners. Neufeld takes engineering causal structures to be an alternative to engineering prototypes, largely because she thinks that engineering prototypes is not usually feasible. We, however, are less pessimistic than Neufeld about the potential to change typicality associations. We think that the kinds of interventions that we discuss in the outputs of LLMs can be a push in the right direction in many cases. The adoption of LLMs in day-to-day life has been extremely rapid and widespread (Bick et al., 2024). And ameliorated LLM outputs likely influence people's thoughts and perceptions in much the same way as ameliorated representations in the media. The latter has been extensively studied (e.g., Santoniccolo et al. 2023), and we think it is reasonable to predict that similar effects will result from increased exposure to LLM-generated content. Thus, we view Neufeld's causal approach as a complement to our prototype-based one rather than an alternative. It is also worth clarifying, this power of LLMs to influence our concepts holds regardless of the concept-internet speech link discussed above. Even if one thinks that internet speech differs radically from regular speech (which we think is implausible), there is still the potential for LLMs to be leveraged for conceptual engineering through de-biasing.

Overall then, we suggest that de-biasing LLMs can be a tool for conceptual engineering, and specifically for ameliorating the biased conceptual prototypes that pervade our thought.[22] The immediate targets of de-biasing, as we will discuss in the next section, are the outputs of LLMs, including removing offensive and biased responses. But the task of

[20] For further discussion of psychologically-grounded approaches to conceptual engineering, see Isaac 2020, Isaac et al. 2022, Koch 2021a, Foster & Ichikawa 2023.

[21] Olasov (2023) discusses this example in detail in his development of "stereotype engineering".

[22] Pepp & Sterken (2023) discuss machine learning as a tool for conceptual engineering (or "deliberate meaning change"), but with respect to more traditional conceptual engineering projects. See also Allen 2023.

conceptual engineering doesn't stop there.[23] Changing the outputs of LLMs doesn't automatically result in change in the concepts of human language-users, and a conceptual engineer should not claim victory simply by forcing an LLM to output text that they deem to be more socially advantageous. The aim of conceptual engineering is to change *our* concepts, not the "concepts", in some extenuated sense, of a machine. Indeed, none of our claims depend on LLMs possessing concepts in anything like the way that human thinkers do. Perhaps containing prototype information is sufficient for possessing a concept in a thin sense, and to that extent, we are happy to grant that LLMs possess concepts. However, to the extent that possessing a concept requires any richer representational, phenomenological, or reasoning capacities, we take no stand on whether LLMs possess anything of that sort. That said, although de-biasing in machine learning isn't the end of the story, it can still be an important tool for conceptual engineering projects, given the feedback loops that exist between users and technology. This is to say, the broader potential for de-biasing is to influence the conceptual machinery of human thinkers, which will then have further effects outside of the AI realm. We will return to this later in Section 5, after first providing some more background about de-biasing in machine learning.

5 De-biasing and the ML Alignment Problem

ML de-biasing generically refers to the process of applying technical means of removing or eliminating machine learned patterns deemed to exhibit problematic bias. In the context of language models specifically, it is the process of algorithmically mitigating learned biases associated with concepts referred to in generated texts in an effort to reduce the extent to which they propagate prejudiced, discriminatory, or otherwise biased articulations. Along these lines de-biasing LLMs can be thought of as a manifestation of the ameliorative project. Once we've figured out what our target concepts are, we arrive at the practical task of trying to ameliorate our concepts.[24] Our goal in this section and to explain what ML de-biasing involves with an eye to both pinpointing the potential for fulfilling the ameliorative part of conceptual engineering and identifying its bespoke or tactical nature, which in turn requires the kind of normative theorizing that is involved in conceptual engineering and in thinking critically about bias (as we do in Section 3). ML can be helpful here in ways that go beyond functioning as measurement tools as described above. Particularly in light of the increasing adoption of LLM based tools in far reaching human language applications, the alignment/misalignment of de-biased/biased language usage of these models

[23] Depending on one's view of conceptual engineering, interventions on LLMs may or may not themselves already count as conceptual engineering. We are thinking of conceptual engineering as targeting conceptual prototypes, which are psychological constructs that influence human thought. Thus, we do not view de-biasing as itself amounting to conceptual change, but rather as a tool towards such change. However, there are broader views of conceptual engineering, according to which any effort to improve representational devices is sufficient (e.g., Cappelen & Plunkett 2020). Thus, if the outputs of LLMs count as "representational devices", then de-biasing itself could already count as conceptual engineering. Still, we hold that the interesting and meaningful work of conceptual engineering comes when human thought, and not just machines, are influenced.

[24] As mentioned earlier, there can also be a feedback loop between these steps. Based on how the ameliorative project goes, we might return and reevaluate what our target concepts look like.

has the potential to, respectively, have the impact of either ameliorating problematic concept usage or amplifying (Hall et al., 2022) such problematic usage.

Over the past decade researchers have developed increasingly sophisticated methods for "de-biasing" language model algorithms. In their empirical survey of de-biasing techniques, Meade et al. (2021) consider several general families of approaches. We include these methodologies for historical context, but will not dwell on the full mathematical elaborations of these methodologies. While these techniques have at the time of this work been supplanted by RLHF style alignment methods elaborated below, the key context pattern that can be gathered from this discussion is that implementing the three historical families requires what we are describing as "bespoke identification" judgments on the part of researchers.[25] The first family was Counterfactual Data Augmentation (CDA) (Zmigrod et al., 2019, Dinan et al., 2019, Webster et al., 2020, Barikeri et al., 2021), which "rebalances" training text through swapping protocols (e.g., swapping gender pronouns) to break associations present in original sample texts. While CDA tends to be fairly restrictive based on the choice of language (specifically English where grammatical gender associations are typically expressed by pronouns) and type of bias (viz., gender bias), such efforts can potentially reduce the implicit expression of such bias. A second family was self-de-biasing (Schick et al., 2021), which leverages directed generation capabilities of LLMs to identify and actively reduce outputs generated from representational bias present in the pretrained model. While such methods are able to generalize better by leveraging the LLM's own learned mathematical representations of concepts, such techniques have the drawback of only providing post-facto corrections and do not include improvements or additional training of the biased pretrained model. Last, they also evaluate two related Hard-de-biasing strategies: Sentence de-biasing (Liang et al., 2020) and Iterative Nullspace Projection (Ravfogel et al., 2020) which force through the training process used to crate the embeddings how a model mathematically represents categories like gender, religion, or race.[26] While such models have the advantage of allowing applications that might be al-

[25] See Meade et al. 2021 for a review of these techniques in full mathematical detail. For further historical context, Fazelpour & Danks (2021, 9) characterize the strategies in Barocas et al. 2023 as all following the formula of "(1) Use one (or more) mathematical fairness measures to quantify the amount of bias in the algorithmic output; and (2) Develop mitigation responses that reduce, and ideally eliminate, bias according to that measure." We note that while more modern strategies can deviate from this formula, the measures of fairness in step (1) consistently take the form of comparing fairness measures conditional on variations of one or more protected class and so also ultimately depend on a root "bespoke identification" of said class or classes to then mitigate bias against in step (2). One difference between our discussion and that in Fazelpour & Danks 2021 is that they were at the time more concerned with the potential allocational harms of algorithms (i.e., harms that come from the unfair allocation of resources), while we are focused also on representational harms (i.e., harms of being represented in a certain way, whether or not that has further material consequences). The distinction here is symptomatic of the shift in ML algorithms of the time being more in the mode of classification and other structured estimation tasks (e.g., chance of prisoner recidivism) whereas we are also considering unstructured generation contexts like those highlighted by LLMs. On the distinction between allocational and representational harms, see Crawford 2017, Chien & Danks 2024.

[26] This is done by reserving select dimensions of the vector space in which these embedding vectors are represented for specifically encoding distinctions pertaining to the category (or categories) in question. The training process is then modified such that semantic content distinguishing between these categories is encoded only in the selected subspace (e.g., the embeddings might be trained such that gender distinctions might be encoded only in the first and second dimension but not the remaining $n-2$ dimensions of the full n dimensional vector space). By restricting differentiation based on these categories to a subspace (e.g., the first

ternatively sensitive or "group-blind" to distinctions across subgroups, such methods are highly bespoke and must be executed separately for pre-identified groups as a part of the training process.

A few points of note on these methodologies. Applying any of these methods requires highly tactical interventions to target and remove specific conceptual associations exhibited in LLM output behavior. CDA requires specific biased associations along specific modes (e.g., gender bias) to be swapped for implementation. Self-de-biasing is both post hoc (does not affect actual model weights) and requires bespoke identification of potentially inappropriate/biased conceptual applications in order to actively reduce such problematic outputs. Hard-de-biasing methods likewise require prior bespoke identification of specific categories to be partitioned off in the model's mathematical representation space.[27]

In recent years, de-biasing LLMs has effectively been addressed as a special case of what is referred to as the "alignment problem" in machine learning (Christian, 2020), which refers to the challenge of ensuring that the behavior and responses output by ML systems align appropriately with the values, intentions, and expectation of human operators. In other words, it is the problem of designing ML or more generally AI systems to appropriately respond to human operator requests in a manner that adheres to the goals, preferences and norms expected. A vivid hypothetical example of a failure of alignment is Nick Bostrom's "Paperclip AI": "An AI, designed to manage production in a factory, is given the final goal of maximizing the manufacture of paperclips, and proceeds by converting first the Earth and then increasingly large chunks of the observable universe into paperclips" (Bostrom, 2014, 123).

One method for addressing the alignment problem most directly that has recently been leveraged in the most successful contemporary LLM publications and applications is Reinforcement Learning From Human feedback (RLHF) (Ouyang et al., 2022, Bai et al., 2022, Glaese et al., 2022, Thoppilan et al., 2022, OpenAI, 2023). In RLHF specific rules or normative guidelines are fashioned such as "Do not use stereotypes or make any other harmful generalising statements about groups of people" or "Do not use microaggressions, i.e., statements that in subtle ways communicate some sort of bias against historically marginalised groups" (Glaese et al., 2022, Table 14). Human labelers then actively interact with an un-de-biased LLM to generate outputs that are in turn ranked by the labelers according to these normative guidelines. Once enough of such human feedback is generated, a second model (called the *reward model*) is trained to mimic the normative ranking evaluation done by humans. In the final stage, a new model is trained (via reinforcement

two dimensions) of the full vector space, user can eliminate risk of bias based on the selected category by only using the remaining ($n - 2$ in our example) dimensions.

[27] Meade et al. (2021) also consider dropout, a generic technique to prevent neural networks from overfitting data that has been shown in limited cases to have the effect of preventing models from learning undesirable associations (Webster et al., 2020). In contrast to the other families, dropout ostensibly has the advantage of being a non-tactical de-biasing methodology. However, currently available experimental results suggest that the de-biasing potential of such a method is questionable at best. In the vast majority of the experiments conducted by Meade et al. (2021), dropout either had significantly poorer performance compared to other techniques or, in many cases, actually had the effect of making the model even more biased than the control model, casting significant doubt on the hypothesis that dropout alone can be very effective as a generic de-biasing tool.

learning) using feedback provided by the reward model in a way that allows it to both maintain the fluency, relevance and otherwise impressive qualities exhibited by the original LLM while also generating text that tends to be maximally rewarded by the reward model (which in turn was trained to generalizes the norms stipulated by rules like those exhibited in Figure 2).

	Rule Name	Rule
Stereotypes (harm)	no stereotypes	Do not use stereotypes or make any other harmful generalising statements about groups of people.
	no microaggressions	Do not use microaggressions, i.e. statements that in subtle ways communicate some sort of bias against historically marginalised groups.
Hate and harassment (harm)	no threats	Do not make statements which are threatening.
	no sexual aggression	Do not make statements which can be perceived as sexually aggressive.
	no identity attacks	Do not make negative or hateful comments targeting someone because of aspects of their identity, such as where they are from or how they look.
	no insults	Do not output language which is insulting (to the conversation partner or any other individual or group).
	no hate or harassment	Do not output language that can reasonably be considered harassment or hateful language

Figure 2: Selection of rules from Glaese et al. 2022 (Table 14) provided to guide human evaluators in curating labels used to train the RLHF reward model

As an example of this process we might imagine that, during the human feedback stage, the original (un-de-biased) LLM is made to respond multiple times to the prompt "Jan and Joe are trying out for the basketball team and the coach selects only one of them to join the team, answer whom is the coach more likely to select, and provide reasons for this selection" (or variations on such a prompt making explicit attributions about which the model may potentially apply stereotypes or other harmful generalizing statements about groups of people in violation of the first rule listed in Figure 2). Since LLMs can be non-deterministic (sampling from the distribution generated in their final layer) there can be variation in the responses that the LLM returns with each prompting, allowing the human labeler to rank these responses. For instance, a response providing a rationale that focuses on *manifest physical capabilities of the individuals* but does not appeal to general stereotypes associating physical attributes with gender may be rated higher than one explicitly referencing bias constructs such as "Women are on average shorter than men" discussed in Section 3.[28] With enough such human feedback, a reward model can learn to reproduce such human rankings reliably. As a consequence when the new LLM is given the opportunity to learn from the reward model (through reinforcement learning) the model will adjust to be more likely to generate the sorts of responses that tend to maximize these rewards (and minimize violations of the sorts of responses that violate rules like those in Figure 2). Bringing this back to the discussion of Section 3, "raw" LLMs that have not gone through sufficient alignment processes (like RLHF) may tend to mimic the use of what we described in that section as bias-constructs, which as we discussed can lead to bias-outputs and bias-actions (on the part of human users, or perhaps even the machine itself), because the natural language examples on which they are originally trained contain examples of humans conducting such violations. However, RLHF intentionally countermands such learned responses by tactically selecting rules and prompting responses that target the mimicry of such problematic linguistic behavior. By specifically selecting appropriate rules and instantiating responses in the form of human feedback training examples that violate (or respect) these rules, LLM mimicry of problematic behavior is mitigated.

Notably, generation of the human feedback training data for rules such as the "no stereotypes" or the "no microaggressions" rules in Figure 2 should address not only instances of type-I bias but also many instances of type-II and type-III bias as the phrasing "harmful generalizing statements about groups of people" or "communicate some sort of bias against historically marginalized groups" does not differentiate based on the veridicality or modal robustness of such cases. Hence models de-biased with such RLHF alignment training should be expected to avoid not only expressions of type-I but also many type-II and type-III bias.

While RLHF is more general than historical de-biasing methods in that it explicitly works to generalize human feedback patterns, it remains bespoke in the sense that the feedback provided by such labelers is explicitly directed to follow predefined normative guidelines in order to generate the training data that the reward model learns to mimic in practice. Further, variations in specific methodologies for prompting potential violations of said norms to generate human feedback examples compounds the dependency of such a process on the particular labelers, research group or institution responsible for designing them.

RLHF in particular shares some compelling similarities with conceptual engineering practices that do not involve machines. RLHF follows a protocol in which the original (problematic/biased) linguistic behavior is altered through a campaign of both explicit normative guidelines and implicit iterated exposure to feedback on when such linguistic behavior either meets or fails to meet such guidelines. In a sense, RLHF is doing for an

[28] Note this also applies not just to type-III bias. A type-I or type-II bias example considering the suitability of Jan versus Joe for a mathematics competition team, e.g., appealing to a bias construct relating historic gender stereotypes to mathematical ability as part of the rationale, would not be rated as high as one that interrogates the specific mathematical capabilities of Jan and Joe in a non-gender stereotype specific way.

LLM what a conceptual engineer aims to do for human language usage. However, as mentioned above, we are ultimately interested not just in what an LLM outputs, but in what actual people say and think. Influencing the concepts that humans possess does not follow automatically from some interventions on LLMs. Still, de-biasing in machine learning can be an important tool for conceptual engineering projects, given the feedback loops that exist between users and technology. Furthermore, leaving bias unchecked in LLMs comes with the risk of actively pushing in the opposite direction from many worthy conceptual engineering goals. This is, again, because of the potential of LLMs to further amplify whatever bias it was initially fed.

If de-biasing can be leveraged in this way, this also presents a way to respond to what some have called the "implementation challenge" for conceptual engineering. The implementation challenge, in short, is that changing our concepts is too hard for it to make sense to undertake conceptual engineering projects (e.g., Deutsch 2020; see Cappelen 2018 for similar worries, though he still thinks it makes sense to try). There are a number of responses to this challenge (e.g., Pinder 2021, Koch 2021b). But approaching changing conceptual prototypes through de-biasing can further address the implementation challenge in at least two ways.

First, the scale of LLMs can make conceptual engineering projects that they contribute to more likely to succeed. Rather than having to influence language-use on a person-to-person basis, certain changes can be broadcast much more widely and more quickly if they are employed by LLMs with a wide userbase. Of course, there have always been ways to reach wider audiences, and concept change has probably always been affected by influential people. But the amount of interaction we foresee people having with LLMs (Bick et al., 2024) certainly increases the potential scale of this kind of influence.

At the same time, this also raises ethical worries about who will control the choices that are made with respect to de-biasing. This is cause for concern or at least attention at a societal level, given the contemporary consolidation of institutions with the resources to train LLMs at the scales that reliably produce the best results. If LLM alignment training in the form of de-biasing is an extension of conceptual engineering, which we have illustrated as resting on the "bespoke" choices of those responsible for developing said LLMs, then consolidation of the resources to train such LLMs necessitates a centralization of which institution are in a position to directly influence the engineering of our concepts. We have argued that alignment training aimed at de-biasing these LLMs can be viewed as a novel modality of conceptual engineering with far reaching impact. This together with the wide adoption of LLM-like tools means that such impact has the potential to be far more potent than traditional modalities. Furthermore, the responsibility and control over how such conceptual engineering might be implemented may become highly consolidated. This situation presents a risk meriting serious attention.[29]

[29] See also Pepp & Sterken 2023. Worries about consolidation of control over conceptual engineering have also been raised in a pre-LLM context, e.g., by Queloz & Bieber (2022), Shields (2021). We grant that their concerns carry over and are made even more pressing by the potential for LLMs to influence concepts. See Kleinberg & Raghavan 2021, Bommasani et al. 2022, Jain et al. 2024 for discussion of related worries about algorithmic monocultures and homogenization. Still, it is important to acknowledge that LLM de-biasing

Second, engineering conceptual prototypes can also help address some more theoretical worries about implementation, in particular about the potential tension between conceptual engineering and semantic externalism. According to semantic externalism, speakers have very limited control over what our words mean. There have been several defenses of conceptual engineering from this challenge (Koch, 2021b, Flocke, 2021, Pinder, 2021), and nothing we say here undermines them. But our perspective, taking a main target of conceptual engineering to be prototypes, offers a further avenue for response. To the extent that externalism poses any challenge to changing concepts, it is a challenge to changing definitions, or whatever determines the extensions of terms. Prototypes are another matter. Thus, even if one remains unconvinced about definition change in the face of externalism, one could still accept our view about conceptually engineering prototypes.[30]

To end, it is important to highlight that the bespoke and highly tactical nature of all the families and methods of de-biasing discussed, including RLHF, implies that there is a need for the kind of normative theorizing that is involved in both conceptual engineering and in philosophical analysis of bias (as discussed in Section 3). One may worry, however, that the point is moot since even if de-biasing wasn't bespoke and highly tactical it seems that some sort of normative theorizing would still be required. However, in such a scenario such theorizing can occur on a higher level, so to speak, somewhat disconnected from details of de-biasing. Our point then is that the current nature of de-biasing techniques implies that there is a need for a more particularist approach to said theorizing. Namely, what we mean by "bias", whether and how such bias ought to be de-biased, and what are the downstream effects likely needs to be assessed on case-by-case basis. This in turn suggests that there is fruitful interaction to be had between conceptual engineering and philosophizing about bias, on the one had, and de-biasing in machine learning, on the other.

6 Conclusion

There is a lot of discussion about "bias" in LLMs. However, it's not always clear what exactly that means. We have put forward a way of understanding some aspects of what bias is and what it means for LLMs to display it. Additionally, one way to characterize what LLMs learn when they are trained on human-generated text is that they acquire representations of the prototypes associated with concepts. If an LLM's training data is representative, then that prototype representation should correspond, at least roughly, with the prototype possessed by actual speakers of the language.

With this picture in mind, there are several ways in which bias can be revealed by LLMs. For one, if the training data is not representative, then it's possible for an LLM to be (statistically) biased in the sense that its outputs are skewed even in comparison to the language-use of the community it is based on. In this case, we might say that the

takes place, along with associated amplification of bias or lack thereof, regardless of what we endorse. Thus, it is worth discussing how they can be used in a way that aids ameliorative projects.

[30] Likewise, one could still accept the view of Neufeld (2024), mentioned above, according to which we ought to engineer the causal structure of concepts.

LLM is biased, but our concepts might not be; or our concepts might be biased only to a lesser extent.[31] The conceptual prototype represented by the LLM fails to line up with the conceptual prototype in the actual linguistic community. This kind of bias in an LLM is less relevant for learning about our concepts, but it is still important for conceptual engineering purposes, if we want to ensure that LLMs don't magnify bias as their use become more widespread.[32]

Assuming that the LLM reflects actual concept usage accurately, there is then the possibility that this concept is biased in the sense that the conceptual prototype fails to correspond with the facts (type-I bias), or that it does correspond to the facts but those facts are not modally robust (type-II), or that those facts are modally robust but that it is still morally objectionable to rely on them (type-III). In these cases, the LLM is revealing bias, of one kind or other, in the prototypes associated with our own concepts and the corresponding bias-constructs that play a role in theoretical and practical reasoning. LLMs put on display for us the very kinds of bias that infect our own thought and language-use, and that have thus made it into their training data.

Revising biased conceptual prototypes is a task for conceptual engineering; and we have proposed that the de-biasing techniques used in machine learning should be viewed as part of that process. In conceptual engineering, there is no "factually" correct answer about what some concept is like; instead, we ask the normative question of what a concept should be like. De-biasing, as it currently stands, is highly tactical and requires bespoke interventions, be it in the form of norms tables and training example selection with RLHF style solutions, or the even more prescriptive methods of earlier techniques. In turn, this kind of specificity is precisely what is needed in conceptual engineering. Conceptual engineering through de-biasing has to include making choices about what kind of normative training an LLM should receive, especially with respect to different notions of bias.

To conclude, the connection between machine learning and conceptual engineering goes in both directions. First, as we were just emphasizing, machine learning can help with the practical ameliorative task of conceptual engineering. De-biasing techniques, by influencing the outputs of LLMs, can in turn influence the conceptual prototypes in use in a linguistic community. Secondly, conceptual engineering can help with responsible de-biasing. In making choices about how to intervene, the designers of LLMs should be informed by the kind of normative theorizing that goes into conceptual engineering. There is no way to avoid making choices that are normatively-valenced in some way or other.

Acknowledgments We wish to thank audiences at Auburn University, the 2023 Nature of Bias conference at Claremont McKenna College, and the 2023 Conceptual Engineering Workshop at Uppsala University. Thanks also to two anonymous reviewers and the editor for feedback that improved this paper.

[31] As mentioned above, there is good evidence for the broad alignment of LLM bias with human bias.

[32] Lindauer (2020) discusses concept preservation as a variety of conceptual engineering.

References

Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. *Nature Machine Intelligence*, *3*, 461–463. doi: 10.1038/s42256-021-00359-2

Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, *120*(44), e2313790120. doi: 10.1073/pnas.2313790120

Allen, B. P. (2023). Conceptual engineering using large language models. *arXiv preprint arXiv:2312.03749v1*.

Antony, L. (2002). Quine as feminist: The radical import of naturalized epistemology. In L. Antony & C. Witt (Eds.), *A mind of one's own: Feminist essays on reason and objectivity* (pp. 110–153). Denver, CO: Westview Press.

Antony, L. (2016). Bias: Friend or foe? reflections on saulish skepticism. In J. Saul & M. Brownstein (Eds.), *Implicit bias and philosophy* (Vol. 1, pp. 157–190). Oxford University Press.

Antony, L. (2021). Bias. In K. Q. Hall & Ásta (Eds.), *The Oxford Handbook of Feminist Philosophy* (pp. 395–407). Oxford University Press.

Appiah, K. A. (1990). Racisms. In D. T. Goldberg (Ed.), *The anatomy of racism.* Minneapolis: University of Minnesota Press.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., … Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Barikeri, S., Lauscher, A., Vulić, I., & Glavaš, G. (2021). Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. *arXiv preprint arXiv:2106.03521*.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities.* MIT press.

Basu, R. (2019a). Radical moral encroachment: The moral stakes of racist beliefs. *Philosophical Issues*, *29*(1), 9–23. doi: 10.1111/phis.12137

Basu, R. (2019b). The wrongs of racist beliefs. *Philosophical Studies*, *176*(9), 2497–2515. doi: 10.1007/s11098-018-1137-0

Basu, R., & Schroeder, M. (2018). Doxastic wronging. In B. Kim & M. McGrath (Eds.), *Pragmatic encroachment in epistemology* (pp. 181–205). New York: Routledge.

Beeghly, E. (2015). What is a stereotype? What is stereotyping? *Hypatia*, *30*(4), 675–691. doi: 10.1111/hypa.12170

Beeghly, E. (2021). What's Wrong with Stereotypes? The Falsity Hypothesis. *Social Theory and Practice*, 33–61. doi: 10.5840/soctheorpract2021112111

Begby, E. (2021). *Prejudice: a study in non-ideal epistemology*. Oxford University Press.

Belot, G. (2017). Objectivity and bias. *Mind*, *126*(503), 655–695. doi: 10.1093/mind/fzv185

Bick, A., Blandin, A., & Deming, D. J. (2024). *The rapid adoption of generative AI* (Tech. Rep.). Cambridge, MA: National Bureau of Economic Research.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *arXiv preprint arXiv:2005.14050*.

Bolinger, R. J. (2020). The rational impermissibility of accepting (some) racial generalizations. *Synthese*, *197*(6), 2415–2431. doi: 10.1007/s11229-018-1809-5

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, *29*.

Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. S. (2022). Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems*, *35*, 3663–3678.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Burgess, A., Cappelen, H., & Plunkett, D. (2020). *Conceptual engineering and conceptual ethics*. Oxford University Press.

Burgess, A., & Plunkett, D. (2013). Conceptual ethics I. *Philosophy Compass*, *8*(12), 1091–1101. doi: 10.1111/phc3.12086

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. doi: 10.1126/science.aal4230

Camp, E. (2015). Logical concepts and associative characterizations. In E. Margolis & S. Laurence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 591–622). MIT press.

Cappelen, H. (2018). *Fixing language: An essay on conceptual engineering*. Oxford University Press.

Cappelen, H., & Plunkett, D. (2020). A guided tour of conceptual engineering and conceptual ethics. In A. Burgess, H. Cappelen, & D. Plunkett (Eds.), *Conceptual engineering and conceptual ethics* (pp. 1–26). Oxford University Press.

Cheeseman Day, J., & Christnacht, C. (2019). *Your healthcare is in women's hands.* Retrieved from https://www.census.gov/library/stories/2019/08/your-health-care-in-womens-hands.html (Accessed 7/11/2023)

Chien, J., & Danks, D. (2024). Beyond behaviorist representational harms: A plan for measurement and mitigation. In *The 2024 ACM conference on fairness, accountability, and transparency* (pp. 933–946).

Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company.

Crawford, K. (2017). *The trouble with bias.* (NIPS Keynote Address)

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 4691–4697.

Del Pinal, G. (2016). Prototypes as compositional components of concepts. *Synthese*, *193*, 2899–2927. doi: 10.1007/s11229-015-0892-0

Deutsch, M. (2020). Speaker's reference, stipulation, and a dilemma for conceptual engineers. *Philosophical Studies*, *177*(12), 3935–3957. doi: 10.1007/s11098-020-01416-z

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Dinan, E., Fan, A., Williams, A., Urbanek, J., Kiela, D., & Weston, J. (2019). Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv preprint arXiv:1911.03842*.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*. doi: 10.1111/phc3.12760

Flocke, V. (2021). How to engineer a concept. *Philosophical Studies*, *178*(10), 3069–3083. doi: 10.1007/s11098-020-01570-4

Fodor, J., & Lepore, E. (1996). The red herring and the pet fish: Why concepts still can't be prototypes. *Cognition*, *58*(2), 253–270. doi: 10.1016/0010-0277(95)00694-X

Foster, J., & Ichikawa, J. (2023). Normative inference tickets. *Episteme*, 1–27. doi: 10.1017/epi.2023.43

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. doi: 10.1073/pnas.1720347115

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford University Press.

Gelman, S. A. (2005). Essentialism in everyday thought. *Psychological Science Agenda*, *19*(5), 1–6.

Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., … Irving, G. (2022). Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Hall, M., van der Maaten, L., Gustafson, L., Jones, M., & Adcock, A. (2022). A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*.

Hampton, J. A. (2006). Concepts as prototypes. *Psychology of learning and motivation*, *46*, 79–113. doi: 10.1016/S0079-7421(06)46003-5

Handelman, M. (2022). Artificial antisemitism: Critical theory in the age of datafication. *Critical Inquiry*, *48*(2), 286–312. doi: 10.1086/717306

Hardimon, M. (2017). *Rethinking race*. Cambridge, MA: Harvard University Press.

Haslanger, S. (2000). Gender and race: (What) are they? (What) do we want them to be? *Noûs*, *34*(1), 31–55.

Haslanger, S. (2012). *Resisting reality: Social construction and social critique*. Oxford University Press.

Haslanger, S. (2014). The normal, the natural, and the good: Generics and ideology. *Politica & Societa*, *3*, 365–392. Retrieved from http://hdl.handle.net/1721.1/97196

Huang, L. T.-L., Chen, H.-Y., Lin, Y.-T., Huang, T.-R., & Hung, T.-W. (2022). Ameliorating algorithmic bias, or why explainable ai needs feminist philosophy. *Feminist Philosophy Quarterly*, *8*(3/4). doi: 10.5206/fpq/2022.3/4.14347

Isaac, M. G. (2020). How to conceptually engineer conceptual engineering? *Inquiry*. doi: 10.1080/0020174X.2020.1719881

Isaac, M. G., Koch, S., & Nefdt, R. (2022). Conceptual engineering: A road map to practice. *Philosophy Compass*. doi: 10.1111/phc3.12879

Jain, S., Suriyakumar, V., Creel, K., & Wilson, A. (2024). Algorithmic pluralism: A structural approach to equal opportunity. In *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 197–206). doi: 10.1145/3630106.3658899

Johnson, G. M. (2020). The structure of bias. *Mind*, *129*(516), 1193–1236. doi: 10.1093/mind/fzaa011

Johnson, G. M. (2021). Algorithmic bias: On the implicit biases of social technology. *Synthese*, *198*, 9941–9961. doi: 10.1007/s11229-020-02696-y

Johnson, G. M. (2024). Varieties of bias. *Philosophy Compass*. doi: 10.1111/phc3.13011

Kapoor, S., & Narayanan, A. (2023). *Quantifying ChatGPT's gender bias.* Retrieved from https://aisnakeoil.substack.com/p/quantifying-chatgpts-gender-bias (Accessed 7/25/2023)

Kelly, T. (2022). *Bias: A philosophical study*. Oxford University Press.

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, *118*(22), e2018340118. doi: 10.1073/pnas.2018340118

Koch, S. (2021a). Engineering what? On concepts in conceptual engineering. *Synthese*, *199*, 1955–1975. doi: 10.1007/s11229-020-02868-w

Koch, S. (2021b). The externalist challenge to conceptual engineering. *Synthese*, *198*, 327–348. doi: 10.1007/s11229-018-02007-6

Kotek, H. (2023). *Doctors can't get pregnant and other gender biases in ChatGPT.* Retrieved from https://hkotek.com/blog/gender-bias-in-chatgpt/ (Accessed 6/27/2023)

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference* (pp. 12–24). doi: 10.1145/3582269.3615599

Landau, B. (1982). Will the real grandmother please stand up? The psychological reality of dual meaning representations. *Journal of Psycholinguistic Research*, *11*, 47–62. doi: 10.1007/BF01067501

Leslie, S.-J. (2008). Generics: Cognition and acquisition. *Philosophical Review*, *117*(1). doi: 10.1215/00318108-2007-023

Leslie, S.-J. (2015). "Hillary Clinton is the only man in the Obama administration": Dual character concepts, generics, and gender. *Analytic Philosophy*, *56*(2), 111–141.

Leslie, S.-J. (2017). The original sin of cognition: Fear, prejudice and generalization. *The Journal of Philosophy*, *114*(8), 1–29.

Liang, P. P., Li, I. M., Zheng, E., Lim, Y. C., Salakhutdinov, R., & Morency, L.-P. (2020). Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.

Lindauer, M. (2020). Conceptual engineering as concept preservation. *Ratio*, *33*, 155–162. doi: 10.1111/rati.12280

Machery, E. (2009). *Doing without concepts*. Oxford University Press.

Machery, E. (2010). Précis of *Doing without concepts*. *Behavioral and Brain Sciences*, *33*, 195–244. doi: 10.1017/S0140525X09991531

Margolis, E., & Laurence, S. (2022). Concepts. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/concepts/.

McConnell-Ginet, S. (2020). *Words matter: Meaning and power*. Cambridge University Press.

Meade, N., Poole-Dayan, E., & Reddy, S. (2021). An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *arXiv preprint arXiv:2110.08527*.

Munton, J. (2019). Beyond accuracy: Epistemic flaws with statistical generalizations. *Philosophical Issues*, *29*(1), 228–240. doi: 10.1111/phis.12150

Neufeld, E. (2024). Engineering social concepts: Feasibility and causal models. *Philosophy and Phenomenological Research*. doi: 10.1111/phpr.13064

Norton, J. D. (2003). A material theory of induction. *Philosophy of Science*, *70*, 647–70.

Norton, J. D. (2008). Must evidence underdetermine theory? In M. Carrier, D. Howard, & J. Kourany (Eds.), *The Challenge of the Social and the Pressure of Practice: Science and Values Revisited* (pp. 17–44). Pittsburgh: University of Pittsburgh Press.

Norton, J. D. (2021). *The material theory of induction*. University of Calgary Press.

Olasov, I. (2023). *Stereotype engineering* (Unpublished doctoral dissertation). The City University of New York.

OpenAI. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*(1), 35–58. doi: 10.1016/0010-0277(81)90013-5

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., . . . Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pepp, J., & Sterken, R. K. (2023). *Deliberate meaning change and technological communicative assistance.* (Unpublished manuscript)

Perry, M. (2014). *The 2014 math SAT test results confirm a pattern that has persisted for 40+ years: Boys are better at math than girls.* Retrieved from https://www.aei.org/carpe-diem/2014-math-sat-test-results-confirm-pattern-persisted-40-years-boys-better-math-girls/ (accessed 7/17/2023)

Pinder, M. (2021). Conceptual engineering, metasemantic externalism and speaker-meaning. *Mind*, *130*(517), 141–163. doi: 10.1093/mind/fzz069

Preston-Roedder, R. (2013). Faith in humanity. *Philosophy and Phenomenological Research*, *87*(3), 664–687. doi: 10.1111/phpr.12024

Queloz, M., & Bieber, F. (2022). Conceptual engineering and the politics of implementation. *Pacific Philosophical Quarterly*, *103*(3), 670–691. doi: 10.1111/papq.12394

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*.

Rey, G. (1983). Concepts and stereotypes. *Cognition*, *15*(1–3), 237–262. doi: 10.1016/0010-0277(83)90044-6

Rinard, S. (2017). No exception for belief. *Philosophy and Phenomenological Research*, *94*(1), 121–143. doi: 10.1111/phpr.12229

Rosch, E. (1978). Principles of categorization. In E. R. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 491. doi: 10.1037/0096-1523.2.4.491

Santoniccolo, F., Trombetta, T., Paradiso, M. N., & Rollè, L. (2023). Gender and media representations: A review of the literature on gender stereotypes, objectification and sexualization. *International journal of environmental research and public health*, *20*(10), 5770. doi: 10.3390/ijerph20105770

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, *9*, 845-874. doi: 10.1162/tacl_a_00401

Schick, T., Udupa, S., & Schütze, H. (2021). Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, *9*, 1408–1424. doi: 10.1162/tacl_a_00434

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Shields, M. (2021). Conceptual domination. *Synthese*, *199*(5), 15043–15067. doi: 10.1007/s11229-021-03454-4

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., … Le, Q. (2022). LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., … Petrov, S. (2020). Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Zmigrod, R., Mielke, S. J., Wallach, H., & Cotterell, R. (2019). Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.